DOCUMENT RESUME

ED 146 173 95 TM 006 473

TITLE	Consortium on Methodology for Aggregating Data in Educational Research. Final Report.
INSTITUTION	Vasquez Associates Ltd., Milwaukee, Wis.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C. Basic Skills Group. Measurement Div.
PUB DATE	31 Jan 77
CONTRACT	NIE-C-74-0123
NOTE	402p.; Reprint from Proceedings of the Conference on Methodology for Aggregating Data in Educational Research (Stanford, California, October 23-24, 1976); Technical Reports 5 and 6 of the original document are copyrighted and therefore not available. They are not included in the pagination. For related documents, see ED 135 813 and 137 317; not available in hard copy due to marginal legibility of original document
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	Classroom Research; Conference Reports; Consortia; *Data Analysis; *Educational Research; *Grouping Procedures; Guidelines; Interdisciplinary Approach; Mathematical Models; Organizations (Groups); *Research Methodology; *Research Problems; School Statistics; *Statistical Analysis; Statistical Bias; Student Characteristics; Teacher Characteristics; Technical Reports; Theories
IDENTIFIERS	*Aggregate Data Analysis

ABSTRACT

The Consortium on Methodology for Aggregating Data in Educational Research was designed as a collaborative effort by research methodologists from economics, education and sociology. Consortium goals were to delineate what was known about the problems associated with the analysis of data from the multiple levels (pupil, teacher/classroom, school) of educational systems, to develop analytical theory to account for these problems, and to organize the findings into guidelines for research and evaluation efforts using aggregated data. Analysis problems that were the focus of the project can be characterized in the following manner: (1) Change in units of analysis/grouping of observations/ecological inference/cross-level inference/aggregation effects, e.g., if one knows the desired level of inference, what complications arise from analyzing data at different levels? (2) Contextual/compositional/structural/"school" effects, e.g., is it important to consider the effects of more than one level of units on the variable of interest? (3) Decisions regarding appropriate units of analysis, e.g., are there general concerns and guidelines for determining the appropriate units of analysis in a given research and evaluation context? Listings of all technical reports, symposia, and conference presentations written, presented or organized under the auspices of the Consortium on Methodology for Aggregating Data in Educational Research are included. (Author/MV)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

Vásquez Associates, Ltd.
P. O. BOX 5630 MILWAUKEE, WISCONSIN 53211

Alma G. Vásquez
President

ED146173

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

FINAL REPORT

NIE C-74-0123

Consortium on Methodology for Aggregating Data in Educational Research

Report to the
National Institute of Education
January 31, 1977

Project Director: Leigh Burstein
University of California, Los Angeles

Co-Principal
Investigator : Michael T. Hannan,
Stanford University

Vasquez Associates, Ltd.
P. O. Box 11530
Shorewood, Wisconsin 53211

TM006 473

Consultants for research, evaluation, and management

ACKNOWLEDGEMENTS

It is impossible to remember and to adequately acknowledge all contributions made by individuals to a project that was in operation for more than 2-1/2 years with the major work being conducted at three different sites.

We wish to thank the many consultants, research assistants and professional colleagues who contributed their time, effort and ideas to the project. Primarily because of them, the process throughout was an enlightening and invaluable professional experience.

We wish to thank the National Institute of Education for providing us the support for what we believed to be an important topic in educational research. The entire staff of the Measurement and Methodology Division of NIE, and especially our Project Officer, Carlyle E. Maw, provided valuable guidance in our effort.

The project would not have been possible without sponsorship of Vasquez Associates, Ltd., Milwaukee, Wisconsin. In particular, we would like to express our gratitude to Dr. George E. Uhlig for his continuing encouragement and cooperation and his willingness to sponsor our research.

Two persons performed vital roles throughout the project without receiving proper acknowledgement of the importance of their work. Ms. Margie Mika, Office Manager, Vasquez Associates, served as both Project Sectetary (until October, 1975) and Bookkeeper. Ms. Antonia Nagy joined the staff as Project Secretary in October, 1975. Margie and Toni had the unenviable task of translating unreadable drafts of highly technical material into polished professional manuscripts. They were also responsible for coordinating communications among the three project locations as well as with NIE. I could not have survived the project without them.

Finally, I wish to thank my co-principal investigator, Michael T. Hannan, for making this whole project possible. My initial interest in the area of data aggregation was stimulated by Mike. His colleagial advice and cooperation and his contributions to the field and to my own thinking are immeasurable.

Leigh Burstein,
Project Director

The Consortium on Methodology for Aggregating Data in Educational Research (Contract NIE-C-74-0123 to Vasquez Associates, Ltd.) was designed to be a collaborative effort by research methodologists from economics, education and sociology (a) to delineate what is known about the problems associated with the analysis of data from the multiple levels (pupil, teacher/classroom; school) of educational systems, (b) to develop analytical theory to account for these problems, and (c) to organize the findings into guidelines for research and evaluation efforts using aggregated data.

The kinds of analysis problems that were the focus of the project work can be characterized in the following manner:

1) Change in Units of Analysis/Grouping of Observations/Ecological Inference/Cross-Level Inference/Aggregation Effects.

THEORETICAL QUESTION -- If one knows the desired level of inference, what complications arise from analyzing data at different levels?

SUBSTANTIVE EXAMPLE -- What are the consequences of analyzing the relation of the mean teacher experience of a school to the mean pupil performance of a school when one wishes to make inferences about how an individual pupil's performance is affected by his teacher's experience?

2) Contextual/Compositional/Structural/"School" Effects.

THEORETICAL QUESTION -- Is it important to consider the effects of more than one level of units on the variable of interest? What are the effects of higher-level independent variables on lower-level dependent variables.

SUBSTANTIVE EXAMPLE -- Is it important to include a measure of academic climate, such as mean family background or mean ability, in models designed to account for individual pupil performance? If so, what are the likely effects of this inclusion?

3) Decisions Regarding Appropriate Units of Analysis.

THEORETICAL QUESTION -- Are there general concerns and guidelines for determining the appropriate units of analysis in a given research and evaluation context?

SUBSTANTIVE EXAMPLE -- How can one determine whether a between-student, between-class, or between-school analysis is appropriate for determining the effects of specific classroom and school characteristics on academic performance?

For the first eighteen months of the project, most of the Consortium time was devoted to examining the effects of analyzing data at a level other than the level at which inferences are to be made; e.g., using grouped data to estimate individual-level relations (Question 1). Project personnel combined earlier findings in economics, education, psychology, sociology and statistics with their own analytical and empirical efforts to develop a theory which specifies the consequences of estimation across levels of analysis in regression-based models. The Consortium work related to change-in-units problems is contained in the January 1975 and July 1975 Progress Reports, in Technical Reports 1-12, 14, 15 and 23-25 and in the symposia and conference presentations (Numbers 2 through 7).

Project work on contextual effects (Question 2) is contained in the July 1975 Progress Report, in Technical Reports 1, 4-6, 13, 15 and 20-22 and in symposia and conference presentations (Numbers 1 and 7).

During the last year of the Consortium's work, an increasing amount of time was devoted to delineating the issues regarding the appropriate units of analysis (Question 3). This question was further considered in Technical Reports 1, 3-6, 8, 10-12, 14-19 and 23 and symposia and conference presentations (Numbers 3 through 6).

To culminate project activities, a Conference on Methodology for

Aggregating Data in Educational Research was organized. On October 23-24, 1976, approximately 25 researchers (see proceeding of Conference on Methodology for Aggregating Data in Educational Research for list of participants) from economics, education, industrial administration, psychology, and sociology took part in a two-day series of panel discussions at Stanford University. Topics of the panels included: the analysis of multilevel data, contextual effects, experimental design considerations, the analysis of qualitative data and estimation from grouped observations in multivariate models. The conference combined papers written specifically for the conference (Technical Reports 16-18, 20-25) with summaries of work done outside the auspices of the contract (presentations by Lee J. Cronbach, Charles N. Henderson, Jr., and Aage Sorenson).

The purposes of the conference were to ascertain the degree of concurrence in thinking about data aggregation in educational research and evaluation and, hopefully, to gain some consensus about avenues for further work. As the papers and the summary of the conference discussion (Technical Report 25) indicate, though there has been much progress in the last few years, substantial differences of opinion remain about the most appropriate methodological models for analyzing multilevel educational data.

The principal investigators feel that substantial progress has been made in the area of change in units of analysis. At the same time, it is becoming increasingly apparent that the work under all three areas addressed by the project (change in the units of analysis, contextual effects, and appropriate units of analysis) are inextricably intertwined elements of the methodology for analyzing multilevel educational data. Future research on issues related to units of analysis in education can be expected to emphasize the development of analytical models to make maximum use of information gathered from all levels of the educational hierarchy. Hopefully, the Proceedings of the Conference on Methodology for Aggregating Data in Educational Research will provide a source document for the direction on further methodological advances in the analysis of multilevel educational data.

Contents of the Final Report

The contents of the final report include listings of all technical reports, symposia, and conference presentations written, presented or organized under the auspices of the Consortium on Methodology for Aggregating Data in Educational Research. Additional documents include revised versions of Technical Report 5 (Journal copy) and Technical Report 6 (in press), and Technical Reports 11 through 25. Technical Reports 15 through 25, as combined, form the Proceedings of the Conference on Methodology for Aggregating Data in Educational Research.

Technical Reports 1 through 4 were submitted as part of the July 24, 1975 Progress Report. The original version of Technical Reports 5 and 6 plus Technical Reports 7 through 10 (Burstein, December 1975; Hannan, April 1976; Poynor, April 1976; and Burstein, April 1976) were submitted to the Project Officers on June 29, 1976.

Products of the Project

The contents of this Final Report, plus all Technical Reports submitted earlier, and the summaries of the symposia and presentations, numbers 1 through 3 in the July 24, 1975 Progress Report, constitute the final products in the fulfillment of Contract NIE-C-74-0123.

Consortium on Methodology for Aggregating Data in Educational Research

Vasquez Associates, Ltd.

Technical Reports under contract NIE-C-74-0123
From the National Institute of Education

No.

1. Burstein, L. "Data Aggregation in Educational Research: Applications." Technical Report No. 1, March, 1975. Also, paper presented at AERA meeting in Washington, D.C., April, 1975.

2. Hannan, M.T., Young, A.A., and Nielsen, F. "Specification Bias Analysis of the Effects of Grouping of Observations in Multiple Regression Models." Technical Report No. 2, April, 1975. Also, paper presented at AERA meeting in Washington, D.C., April, 1975.

3. Hannan, M.T., Freeman, J.H. and Meyer, J.W. "Specification of Models for Organizational Effects: A Comment on Bidwell and Kasarda." Technical Report No. 3, May, 1975. Also, published in American Sociological Review, 1976, 41: 136-143.

4. Burstein, L. and Knapp, T.R. "The Unit of Analysis in Educational Research." Technical Report No. 4, August, 1975.

5. Burstein, L. "The Choice of Unit of Analysis in the Investigation of School Effects: IEA in New Zealand." Technical Report No. 5, October, 1975. Also published in New Zealand Journal of Educational Studies, May, 1976.

6. Burstein, L. and Smith, I.D. "Choosing the Appropriate Unit for Investigating School Effects: The IEA Survey of Science Education in Australia." Technical Report No. 6, Revised, October, 1976. To be published in Australian Journal of Education, 1976.

7. Burstein, L. "The Use of Data from Groups for Inference About Individuals in Educational Research.", Technical Report No. 7, December, 1975. Also, unpublished Ph.D. Dissertation, Stanford University, December, 1975.

8. Hannan, M.T. "Aggregation Gain Reconsidered." Technical Report No. 8, April, 1976. Also, paper presented at AERA meeting in San Francisco, April, 1976.

9. Poynor, H. "Spurious Aggregation and the Units of Analysis." Technical Report No. 9, April, 1976. Also, paper presented at AERA meeting in San Francisco, April, 1976.

10. Burstein, L. "Assessing Differences between Grouped and Individual-Level Regression Coefficients." Technical Report No. 10, April, 1976. Also, paper presented at AERA meeting in San Francisco, April, 1976.

11. Hannan, M.T. and Freeman, John. "Obstacles to the Comparative Study of Organizational Effectiveness." Technical Report No. 11, June, 1976. Also, paper presented at Symposium on Organizational Effectiveness, Carnegie-Mellon University, June, 1976.

12. Glendening, L.K. "The Effects of Correlated Units of Analysis: Choosing the Appropriate Unit." Technical Report No. 12, September, 1976. Also, paper presented at AERA meeting in San Francisco, April, 1976.

13. Meyer, John W. "The Effects of Education as an Institution." Technical Report No. 13, August, 1976.

14. Hannan, M.T. "Methodological Problems in Analyzing the Effects of Educational Organizations on Social Structure." Technical Report No. 14, August, 1976. Also, paper presented at conference on Polish and American Youth: Higher Education and Changes in Social Structure, Warsaw, August, 1976.

The following papers were presented at the Conference on Methodology for Aggregating Data in Educational Research at Stanford University, October 23-24, 1976 and are incorporated here as Technical Reports Nos. 15 through 25.

15. Webb, N., Capell, F. and Nielsen, F. "Summary of Presentation and Discussion, Conference on Methodology for Aggregating Data in Educational Research," Stanford University, October 23-24, 1976. Technical Report No. 15, October, 1976.

16. Burstein, L. and Linn, R.L. "Detecting the Effects of Education in the Analysis of Multilevel Data: The Problem of Heterogeneous Within-Group Regressions." Technical Report No. 16, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

17. Hannan, M.T. and Young, A.A. "On Certain Similarities in the Analysis of Multi-Wave Panels and Multilevel Cross-Sections." Technical Report No. 17, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

18. Keesling, J. W. "Components of Variance Models in Multilevel Analysis." Technical Report No. 18, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

19. Henderson, Charles N. Jr. "Comments on 'Statistical Analysis of a Complex Model Involving Fixed and Random Factors, Unequal Subclass Numbers, and Higher Level Covariates'" Technical Report No. 19, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

2 Alwin, D. F. "Structural Equations Models of Contextual Effects." Technical Report No. 20, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

21. Meyer, J. "Research Designs Required for Contextual and Aggregate Studies of School and Schooling Effects." Technical Report No. 21, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

22. Young, A.A. and Erbring, L. "Unraveling Classroom Effects: A Multilevel Feedback Model of Contextual Effects." Technical Report No. 22, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

23. Glendening, L. and Porter, A.C. "Independence and Selecting the Appropriate Unit of Analysis (or the Appropriate Analytic Model)." Technical Report No. 23, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

24. Hannan, M.T. and Young, A.A. "Small Sample Results on Estimation from Grouped Observations." Technical Report No. 24, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

25 *Maw, C.E. "Problems of Data Aggregation and Cross-Level Inference with Categorical Data." Technical Report No. 25, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

*An earlier version of this paper was presented at the AERA annual meeting in San Francisco, April, 1976.

Consortium on Methodology for Aggregating
Data in Educational Research

Vasquez Associates, Ltd.

Symposia, Paper Presentations, and Conferences under Contract #NIE-C-74-0123 from the National Institute of Education.

1) "The Unit of Observation, The Unit of Analysis, and The Independence of Observations in Educational Research," 1975 "Upstate New York" Measurement and Statistics Meeting, Rochester, New York, May 1975. (Leigh Burstein).

2) Experimental Session on Data Aggregation in Educational Research: Applications and Recent Developments, American Educational Research Association Annual Meeting, Washington, D.C. 1975. (Participants: Leigh Burstein, Michael T. Hannan and Carlyle E. Maw)

3) Symposium on Contextual Effects in Education, American Educational Research Association Annual Meeting, Washington, D.C. 1975. (Participants: Duane Alwin, Leigh Burstein, Maureen Hallinan, Robert M. Hauser, John Meyer and David E. Wiley)

4) The Unit of Analysis and Data Aggregation in Educational Research, American Educational Research Association Annual Meeting, San Francisco, California, April 1975. (Participants: Leigh Burstein, Linda Glendening, Michael T. Hannan, J. Ward Keesling, Robert L. Linn, Andrew Porter and Hugh E. Poynor)

5) "Obstacles to the Comparative Study of Organizational Effectiveness," Symposium on Organizational Effectiveness, Carnegie-Mellon University, Pittsburgh, June 1976. (Michael T. Hannan and John H. Freeman)

6) "Methodological Problems in Analyzing the Effects of Educational Organizations on Social Structure." Conference on Polish and American Youth: Higher Education and Changes in Social Structure, Warsaw, Poland, August 1976. (Michael T. Hannan)

7) Conference on Methodology for Aggregating Data in Educational Research, Stanford University, Stanford, California, October 23-24, 1976 (See Conference Proceedings for list of participants.)

TECHNICAL REPORTS

Technical Reports No. 5 and 6 were removed from this document prior to being submitted to the ERIC Document Reproduction Service for microfilming.

AVAILABILITY:

Burstein L. "The Choice of Unit of Analysis in the Investigation of School Effects: IEA in New Zealand." Technical Report No. 5, October, 1975. Also published in New Zealand Journal of Educational Studies, May, 1976.

Burstein, L. and Smith, I.D. "Choosing the Appropriate Unit for Investigating School Effects: The IEA Survey of Science Education in Australia." Technical Report No. 6, Revised, October, 1976. To be published in Australian Journal of Education, 1976.

Obstacles to the Comparative Study of Organizational Effectiveness*

Michael T. Hannan
Stanford University

John H. Freeman
University of California, Berkeley

Technical Report No. 11

June, 1976

Consortium on Methodology for Aggregating
Data in Educational Research

---

*This research was conducted under National Institute of Education Contract #NIE-C-74-0123

Vasquez Associates, Ltd.
1744 N. Farwell Ave.
Milwaukee, Wisconsin 53202

OBSTACLES TO THE COMPARATIVE STUDY OF ORGANIZATIONAL EFFECTIVENESS[*]

Michael T. Hannan
Stanford University

John Freeman
University of California, Berkeley

Presented at Symposium on Organizational Effectiveness,
Graduate School of Industrial Administration, Carnegie-Mellon University, June, 1976.

[*] This research was supported by National Institute of Education Contract NIE-C-74-0123

OBSTACLES TO THE COMPARATIVE STUDY OF ORGANIZATIONAL EFFECTIVENESS

Effectiveness is one of the strongest and most persistent themes in the literature on organizations. Most theories of organizations introduce effectiveness considerations and many research reports comparing organizations claim to speak to effectiveness issues (see Price 1968 for a literature review). Despite the considerable activity there is little evidence of any cumulation of knowledge on the relationship of organizational characteristics to effectiveness. The lack of cumulation may be attributed to theoretical inadequacies, to methodological deficiences in published research, or to both (cf. Steers 1975).

We agree that the quality of both theory and methodology contribute to the lack of cumulative character in the literature on comparative effectiveness. Below we treat a number of specific problems of each type. The thrust of our argument, however, is to point to a more fundamental problem with attempts at comparing effectiveness across organizations. Most definitions of organizational effectiveness introduce some normative standard against which to evaluate organizational performance. Most often the standard is the set of goals of each organization included in the comparison. Sometimes the standard is the researcher's view on what behavior is appropriate. The two different sets of standards introduce different conceptual obstacles to the comparative study of effectiveness.

When the standard against which performance is evaluated is the set of goals of the organization, comparative studies are hampered by logical difficulties similar to those that arise in attempts at making inter-personal comparisons of utilities. We argue that current conceptions of organizational goals do not permit the establishment of organizational preference functions that have strong enough properties to justify inter-organizational comparisons of goal-attainment.

The second set of standards are those of the researcher (or critic in some cases). As long as the researcher employs his standards consistently there is no obstacle systematic comparisons of (his) goal-attainment across organizations. The status of such comparisons is very different. Whether or not it is made explicit, they are normative. Such comparisons depend intimately on the peculiarities of the observor and do not have inter-observor agreement. Obviously such comparisons do not lend themselves to falsifiability. In short, they are not oriented to scientific activity but to normative or engineering ends.

The literature on organizations suffers from a somewhat general failure to distinguish engineering activity from scientific activity. The problem is as old as the Scientific Management tradition. The goal of Scientific Management was an obvious form of social engineering. Yet the writings on the subject attempted to cumulate the engineering experience into general laws about organizations. One does not have to claim that evidence from applied studies has no relevance to basic science to object to an uncritical shifting from engineering to scientific orientations.

The very term effectiveness suggests an engineering focus. We suppose that materials engineers, say, charged with the obligation of certifying whether or not a proposed structure will withstand certain expected stresses will likely formulate his conclusions in terms of the effectivenss of various alloys for the design. Likewise a construction engineer may debate the effectiveness of various designs. On the other hand, no solid state physicist attempts to use terms like effectiveness in formulating and testing laws that explain the behavior of substances under stress. To do so would be to introduce theoretically irrelevant issues (goals) into the theory.

Our general theme is that effectiveness is a term from the language of applications and engineering but not of abstract scientific theory and research. To introduce effectiveness considerations into attempts at formulating and testing general laws about organizational behavior confuses the two realms of activity to the detriment of both engineering and science. We have no objections to formulating organizational design issues as they impact on a particular organization in terms of effectiveness. All of our objections apply to attempts at comparing organizations in terms of effectiveness when the comparisons are treated as tests of general theories of organizations.

Two Perspectives on Effectiveness

The term effectiveness is used in a bewildering variety of ways in the organizations literature (Steers 1975). Most of these can be seen to reflect one of two distinct emphases: survival or goal-attainment. From the first perspective the organization is effective if it manages to maintain an inflow of essential resources from its environment (Yuchtman and Seashore 1967). From this perspective the effectiveness problem concerns relations with the environment and particularly managing the environment (Pfeffer 1976; Hirsch 1975). The more classical usage concerns goal-attainment. An organization is effective if it meets or surpasses its goals. Most commonly the goals considered are the public goals of the organization. Sometimes the term effectiveness is used with reference to the organization's meeting the goals of leading coalitions or of members generally (Simon 1957).

It is important to keep these two usages separate. They refer to different organizational processes and both the strategies and structures involved may be quite different. Neither one is a surrogate for the other. Nothing we know about organizations suggests that organizations meeting public goals are more likely to survive or that surviving organizations are more likely to meet their publicly stated goals. Unless one can argue convincingly to the contrary, he should not slip indiscriminately from one meaning of the term to the other.

We will follow the more normal usage and use the term effectiveness to refer to goal-attainment (Etzioni 1964: 8). Survival (in an environment) seems a satisfactory term to describe the second meaning of the term effectiveness.

There is a third common meaning of effectiveness that deserves mention: adaptation (Steers (1975) finds that many discussions of effectiveness use the imagery of adaptation). In one sense adaptation is not conceptually different from survival: an organization is adaptive to the extent that it can adjust to changing environmental circumstances. In a second sense, adaptation can be seen as a dynamic aspect of goal-attainment: an organization is adaptive to the extent to which it can adjust output performance to meet goals under varieties of environmental circumstances. Below we discuss the need for dynamic perspectives in the study of organizational performance. Beyond that we see no need for a separate discussion of adaptation-effectiveness.

We have elsewhere (Hannan and Freeman 1974, 1976) discussed the usefulness of modeling selection (non-survival) in populations of organizations. We have also dealt with the methodological problems that arise in such a study (Freeman 1975). Therefore, we concentrate more on obstacles to the comparative analysis of (goal-attainment) effectiveness. This analysis is followed by a brief discussion of issues in the comparative analysis of survival.

Effectiveness and Efficiency

Within the tradition that emphasizes goal-attainment, effectiveness is distinguished from efficiency. There is widespread agreement that the former refers to goal-attainment while the latter refers to the costs incurred in goal-attainment (usually unit cost per output). That is, effectiveness considerations are not made conditional on resources committed and used while efficiency introduces cost comparisons.

Having raised the question of cost comparisons, we are led to ask whether an unconditional analysis of goal-attainment has much meaning. Don't all performance comparisons use at least implicit cost comparisons? The question reduces to the relevance of resources and cost constraints. Organizations that do not face resource constraints can presumably choose to stress either effectiveness or efficiency. Those that employ finite resources (either because they operate in limited environments or they face competition) do not have this luxury. When resources are scarce, performance evaluations that do not take costs into account are not very useful.

Organizations that do not face resource constraints must be very rare. Only for such organizations do we profit by distinguishing efficiency from effectiveness. Consequently, we have chosen not to distinguish the two. We use the term effectiveness to refer to goal-attainment, but insist at a number of places in our argument that resource constraints must be considered.

CONCEPTUAL AND METHODOLOGICAL PROBLEMS WITH GOAL-EFFECTIVENESS

In its most common usage, effectiveness refers to the degree of congruence between organizational goals and organizational performance. Effectiveness is well defined only if both goals and performance are well defined and the comparison of the two is meaningful. Our position briefly stated is as follows. Organizational goals are both difficult to conceptualize and difficult to measure. Performance, although conceptually simpler, is also difficult to measure. Nonetheless, there is no insuperable difficulty in utilizing either concept in developing logically consistent and falsifiable propositions concerning organizational structure and behavior. The possibility that both goals and performances may be defined and measured adequately does not guarantee that attempts at measuring correspondences between the two will yield a scientifically useful concept. We allow that such goal comparisons can be meaningful when applied to a single organization whose goals are well understood. This may be so in engineering applications where the analyst has access to organizational goal makers. On the other hand, we claim that these comparisons are not meaningful across organizations and that propositions concerning comparative effectiveness are untestable.

A. Goals

A great many theorists have pointed to the difficulties involved in defining organizational goals (Simon 1964; Etzioni 1960). In light of these many complications, some of which discussed below, one is tempted to formulate theories of organizations that do not employ the concept. Such a resolution to the problem is not satisfactory. Social scientists have distinguished formal organizations from families, communities, and other forms of social organization precisely because they have specific and limited goals (Blau and Scott 1962). Were we to drop goals from consideration, there would be no need for special theories of formal organizational structure and behavior.

need transition here

Simon (1964) defines goals as "value premises that serve as inputs to decision-making" and also equates goals with preferences. Goals are the values or preferences that concern the relation of the organization to to some criterion or set of criteria. Goals define both the criteria that are relevant and a preference ordering (at least) of outcomes on each criterion.

A The major conceptual difficulties arise because of the likely multiplicity of organizational goals. While some organizations have single public goals (the firm is the classic example), many others have multiple, ostensible, publically legitimated goals. Prominent examples include universities (research and teaching), prisons and public mental hospitals (custody and rehabilitation), and tertiary care hospitals (primary medical care, medical education, and research). Virtually all public agencies and bureaus have very many public goals. For example, the number and diversity of goals of agencies like HEW and the Commerce Department boggle the imagination.

A. Many, if not all, organizations have private goals (sometimes called operative goals) that diverge to some extent from public goals. It is worth distinguishing several important cases:

a) Dominant coalitions often develop the goal of preserving their position of power at almost any cost to the organization. Often this leads them to pursue very cautious organizational strategies (even when the public goals call for decisive action). The classic analysis of this phenomenon is Michels' (1949) account of the subversion of the public goals of European socialist unions and parties by dominant coalitions intent on avoiding risky action. Similarly, Galbraith (1967) argues that the dominant coalitions (called the technostructure) in modern concentrated industries subvert the goal of profit maximization with its attendant risks. Instead they pursue policies of steady (and presumably safer) growth. Baumol (1959) agrues that managing coalitions privately maximize sales and Williamson (1964) claims the operative goal is the maximization of executive salaries (by such strategies as maximizing size of organization). In none of these cases does the dominant coalition announce a change in goals. But insofar as these groups control organizational action, they lead the organization to conform to goals that are different from the public goals. Performances that may appear quite unsatisfactory relative to public goals may be quite satisfactory from the perspective of what have become operative goals.

b) Subunits of organizations often develop their own goals. The process may be identical to that just discussed. Leaders of every subunit may subvert organizational goals in exactly the manner in which dominant coalitions do.

c) Members also bring their own goals to organizations and the degree of correlation between public and private goals is problematical. A number of theorists have argued that problems created by this disjuncture depend systematically on such variables as the "inducements" (both quality and kind) provided by the organization, the "contributions" members are called upon to provide, and the alternatives available through withdrawal (cf. Simon, 1957, 1964; March and Simon, 1958). Their analyses lend little comfort to those who claim that organizations (even firms) have simple goals.

4. Many organizations have *non-specific goals*. Consider the university's goal of advancing the store of useful knowledge or a police agency's goal of protecting the public safety. A very great range of more specific goals may

be seen (by at least come relevant groups) as consistent with the public goals. As a consequence the goals may differ from unit to unit. Moreover, over time any single unit may emphasize one or another of the set of goals consistent with the public goal, depending on the nature of the practical contingencies facing the organization. So even though no set of participants may be subverting public goals, each set may follow very different operative goals.

A second broad problem concerns the temporal dimension of goals. Just as individuals may differ in their time horizons or time discounts (their preferences for immediate versus long-term gratification), so too may organizations. Public goal statements rarely if ever give a clue as to the relevant temporal considerations. Should we consider the short-run or the long-run or both? The many published empirical studies that employ cross-sectional data on samples of organizations (see for example, Lawrence and Lorsch, 1969, Seashore et al., 1960) tacitly take the short-run perspective. Whether or not this is appropriate depends on the nature of the goals function for each organization. To the extent that the goals function stresses quick return on investment (as in many business ventures, disaster relief organizations, military field units, etc.) the short-run outcomes should be given highest weight. For those organizations that orient more towards continued production (e.g. many other types of business ventures, universities, research and development organizations, etc.) the year to year fluctations in performance should be discounted and the average performance over longer periods emphasized.

The conceptual problem is that we do not know how organizations discount time. Two organizations with the same goals operating with the same structure in the same environment may place a very different emphasis on speed of return on investment. One organization may capitalize on some situations in a way that increases both the probability of quick favorable outcomes and the risk of long term decline. The other may eschew the quick return in favor of long run security. As we point out below in such cases both short term and long term comparisons of performance between the two are misleading effectiveness comparisons. The general point is that it is not enough to know the output criteria in the goals function. The analyst must as well know the premium placed on speed of return.

We would not be at all surprised to find that different time perspectives are employed at different levels within organizations. In fact Barnard (1938) has argued that there is a positive relationship between the length of the time perspective employed by a unit and its place (height) in the organizational hierarchy. If so this produces an interesting problem. Consider an organization with a single known goal. If this organization has several layers in its hierarchy which employ quite different time perspectives, with respect to the timing of desired outcomes the organization has a continuum of goals. Unless we are willing to take some kind of average of the time perspectives of all the units as the organizations time perspective, there is no longer any simple goal.

Each of the issues we have identified point to the subtlety of the concept of organizational goals. Each issue also poses quite ~~serious~~ challenges to attempts to measure organizational goals. In particular, we see that organizational goals are ① multi-dimensional, ② often unobserved and ③ temporally indeterminate. Due to the non-specificity of so many goals, it will often be unclear exactly what organizational action or what dimensions of outputs are relevant to goal-attainment. Careful research design could in principle overcome each of the seeming problems with the measurement of goals. But the measurements that would result would be complex (they would consist at a minimum of vectors of value inputs). More important they would, we believe, provide only very rough preference orderings of outcomes.

It is appropriate, given the numerous problems with the concept and with its measurement, to reopen the discussion of its role in organizations theory. Recall that we defined goals as preferences. This suggests that we consider the use ~~role~~ of preferences in micro-economics, where they play a fundamental role. All consumer theory (including the logically equivalent theory of the firm) depends intimately on individual preferences. The entire causal role is played by preferences and the actions of the environment (summarized in prices). But preferences are not measured directly. Their role (more precisely, the role of hypotheses regarding preference functions or indifference curves) is to permit the formulation of testable hypotheses relating prices and consumer behavior.

Many of the problems we have listed are also problems in the conceptualization and measurement of individual preferences. Our point is that these difficulties do not impede the use of unmeasured preferences in the formulation of theories with strong empirical consequences. Why not the same role for organizational goals? Goals and environmental configurations together causally determine organizational behavior. Following the lead of micro-economics we suggest treating goals as unmeasured causal variables and to use propositions involving goals to derive falsifiable propositions relating environmental characteristics and organizational action.

B. Organizational Performance

The second concept involved in effectiveness is organization performance--usually output. In this case the problems appear more methodological than conceptual. Therefore we stress the former issues as they arise in comparative studies of the relationship of variations in structure (or other characteristics of organizations) on outputs.

1. Time Perspective

The problem just discussed is also problematic when we consider performance. In changeable environments organizations are faced with a strategy problem involving specialism or generalism. When the environment shifts unpredictably among a series of states to which any organizational structure is differentially well adapted, organizations must choose to either adopt a structure that does well in one or a limited number of environmental states, that is to specialize to certain environmental outcomes, or to adopt a more generalized structure that is not particularly well suited to any single outcome but which does passably well in all or most of them. We have discussed the manner in which the form of environmental variation affects the optimal strategies for organizations elsewhere (Hannan and Freeman 1974; 1976). Two issues from that discussion are relevant here.

The typical generalist strategy involves the accumulation of varieties of excess capacities (e.g. stockpiling raw materials and deversifying product lines by manufacturing firms; perparation for "brushfire" and nuclear war by armies). Whatever state the environment adopts at any instant, the generalist will appear maladapted. All of the excess capacities that are useful in other environmental outcomes will appear wasteful in the present environment. In contrast, any organizations that specialize in the environment prevailing at that moment will appear adaptive. That is, performance given available resources will be high relative to all other organizations in the field.

On the other hand, if the population ecology theory we advanced is true,[1] we can find environments where over time populations of generalists will outperform populations of specialists. This possibility allows that two investigators studying the same collection of organizations may draw exactly opposite conclusions from cross-sectional and time series performance comparison. On the average specialists in their optimal environments will outperform the generalists in every cross-section. Over time as long as the environment is changing the generalists will tend to outperform the specialists. The situation is actually more complicated than this since the specialists who do not hit their optimal environment will do even worse than generalists. Consequently, in any cross-section there should be greater variance in performance among populations of specialists than generalists. We obviously cannot deal with these complexities in the usual cross-sectional design. With longitudinal data, these considerations lead to testable hypotheses.

2. Bounding Systems

One of the most difficult problems in testing propositions relating structure to performance is to isolate systems in which the effects of any element of structure can be distinguished from the effects of other variables (including other organizational structures). The problem of course is not unique to organizational analysis, but is endemic to all empirical scientific activity (Blalock 1964). Nonetheless, the problem hits with considerable force in this area.

At a minimum one must be able to identify precisely the structures that are involved in any output performance. Organizational performance can often be decomposed into a series of discrete actions where each action can be identified with a particular actor or subunit. For each such action one can identify a relevant organizational structure. A description of an organization in these terms will be a very microscopic description of structure. Alternatively, one can identify the macroscopic aspects of structure that impinge on all action within the organization. The question arises: should one employ a microscopic or macroscopic perspective on structure in evaluating performance outcomes? Both strategies find their advocates in the literature. For example, Scott et al. (1972) argue for studying structure as close to task performance as possible. Bidwell and Kassarda (1974) argue (for school organizations) that one should focus on the highest levels of organization, the levels at which policy decisions that affect the entire organization are made.

One cannot decide between these alternatives without considering the nature of the process under study. It is obvious that a microscopic perspective yields more detail and allows a finer specification of the manner in which structure enhances or impedes task activity. On the other hand, we should not expect that a series of microscopic perspectives taken together will provide as clear a view on the overall relation of a structure to performance in any organization than will a more macroscopic perspective.

Absence of hierarchy (in Simon's 1962 sense of a partial ordering of units as in a set of Chinese boxes) also complicates the system closure problem. The more horizontal links one finds in organizations the less clear at the microscopic level will it be what structures and what resources were activated in any performance.

Our interest is in the different system closure demands that the alternative strategies impose on the researcher. It seems that the more finely one divides the organization the more difficulty one will have in isolating exactly what fraction of effort (manpower and other resources) are devoted to the task in question as opposed to other tasks. To this extent one will have difficulty in isolating the effects of elements of structure from resources. In this sense we propose that one will face a tradeoff in a microscopic study between the fineness of detail on the causal structure and the difficulty in isolating the system. ①

Complexity of the performance also increases the difficulty in isolating the structures and resources that came to bear on the performance. ② Ordinarily the more complex the product the more joint is the action that produced it. Insofar as joint action extends across small scale task organization lines, it will be difficult both to identify which structures are involved and to determine what levels of effort were expended in the performance. Any particular subunit structure may be given undue credit or blame for the contributions of joint actors.

Many of these difficulties can be avoided by considering the systems to be entire production processes and all of the units involved in the process. Often the production process will involve all units in an organization and in other cases a single organization will contain many separable production processes. Presumably one would distinguish boundaries among production processes by examining the gradients in flows in information, personnel, and materials among positions and units. It is important to notice that a production process often extends beyond the boundaries of legally constituted organizations. Most specialized organizations can be considered nodes in a large-scale division of labor. Just as it is difficult to draw meaningful boundaries among subunits of a single organization, it may be difficult to do so in communities or networks of organizations (Hawley 1950; 1968). The production of basic research provides a clear illustration of the problem. Since so much scientific knowledge is communicated rapidly through the research community, discoveries in any research institute may pay off as rapidly in other institutes (in terms of improved performance in making future breakthroughs) as in the institute that made the discovery. To the extent that information flows freely across institute boundaries it appears as though the science as an institution is a more meaningful production system than is the staff and facilities of any research institute. The more general point, then, is that the more permeable the boundaries of any organization (or subunit of an organization) to flows of information, personnel, and other strategic production resources, the more difficult it is to close the system at that level and the more meaningful it is to formulate and test propositions relating structure to performance at some wider system level.

(5)

tautology

3. Levels of Analysis

A related set of issues arise in selecting a system level at which to test propositions. Should we consider the properties of outputs as they "leave" the organization? Should we consider the organizations impact on the larger system? Does it suffice to employ information on average levels of output quality or must we utilize information on distributions? The answers to these questions depend on the form of the propositions under study, the nature of the organization, and the relationship between properties of the inputs (objects worked on) and eventual outputs.

To attempt a complete analysis of these issues would require much more space than we have here. It will help compress the discussion to consider a single type of organization: schools. Suppose that the only output of schools is instruction in some subject matter; and that measurements of knowledge of the subject is termed academic achievement. Suppose ~~finally~~ also that we believe that properties of students such as ability, socio-economic background and aspirations are important determinants of achievement. Consider the problem of testing hypotheses that relate school or school district structure to achievement. Most research that takes this form (the "school effects" literature) relates individual pupil achievement to his/her relevant characteristics and to the properties of school structure. For example, one might regress each students achievement on his/her IQ, race, parent's SES, and on the proportion of school professional work force that is not involved primarily in teaching.

Bidwell and Kassarda (1974) criticized the school effects literature for choosing the wrong level of analysis: school instead of school district. They point out that the structures that regulate the flows of essential inputs and that buffer the technological core from the environment are district level structures. They proceeded to use measures of district structural features and to study district average achievement rather than individual outcomes. They argued that the choice of level of system at which to measure the structural variables settles the choice of level at which to measure outcomes.

In no sense does choice of system level at which to conceptualize and measure structural variables determine the level at which one is to conceptualize and measure outcomes. Rather the choice of each should be determined by the nature of the causal process under study. If the causal process is stated in terms of the effects of structure on the careers of individual students (as was the case with Bidwell and Kassard's paper), the pair:

[district structure and individual outcome] is appropriate. We argued ---- -- Suppose instead that one were studying the hypothesis that a certain structure alters the correlation between ability and achievement (perhaps the structure was advertised to have such a consequence and was introduced for that reason). Could one study this process utilizing average outcomes over students? No. A test of such a proposition requires distributions of outcomes within structures (districts, in this example).

We have critized their research on the grounds that the level of aggregation of outcomes is not consistent with the nature of the causal processes proposed. Bidwell and Kassarda (1976) have responded to our criticism by noting that aggregate events are summaries of individual events and that consequently the investigator may choose any level of aggregation according to his "design".[2] No one will dispute that a distribution consists of information about the properties of individual events. But it is hard to imagine someone defending the view that because of this logical dependence one may shift freely from considering the distribution to considering its center (the average).

When are distributional issues relevant to tests of propositions about structural effects on performance? The key to the issue is the extent to which the investigator expects properties of the inputs (children in this case) to have impacts on outputs and the likelihood that the distributions of the relevant properties differ among organizations. If no individual properties are relevant, analyzing averages will yield the same results as analyzing individual outcomes. Further, if all of the relevant individual properties are controlled, grouping observations by district will not systematically alter results (Hannan, Young and Nielsen 1975). But, if some individual properties relevant to output are left uncontrolled, one can expect considerable divergence between analyses of individual outcomes (as in most sociological school effects studies) and of average outcomes (as in Bidwell and Kassarda) (see Hannan, Freeman and Meyer 1976 for substantive evidence).[3]

This case, choosing between the entire distribution and its average is the most subtle of the levels of analysis issues since the language used to describe the results in each case are so similar. We have no such difficulty discovering the data requirements to test hypotheses about effects on variances, correlations, etc.

Performance cannot always be measured internally to the organization. For example, to evaluate the performance of public health organizations in preventive medicine it is not sufficient to observe the actions of members of those organizations or the health status of persons they contact. Rather one must evaluate some measures of health status in a target population. That is, for some types of organizations the unit within which to evaluate the outcome is some population much larger than the population that has direct experience with the organization. In such cases the number of uncontrolled factors that operate are presumably very large and it is extremely difficult to distinguish the effects of organizational characteristics on the target population from properties of the population and uncontrolled effects of the environment.

4. The Problem of Endogeneity

To this point we have tacitly accepted the classical view of modeling performance (that also is incorporated into models of effectiveness). Performance is equated with the quantity and quality of outputs. Characteristics of the environment, the organization's technology, the quality of inputs are considered given fixed constraints facing the organization. It is time to consider the appropriateness of these assumptions.

Why should the qualities of inputs be taken as a fixed constraint in modeling performance? Many organizations have control over their inputs. Organizations may do well or poorly in identifying appropriate inputs and screening out those that do not fit the organization's needs. It seems reasonable to consider this an aspect of performance (Aldrich and Pfeffer, 1976). In fact, Yuchtman and Seashore consider this the most useful defining criterion of effectiveness (1967). Consider the action of elite private schools. Such schools may have high levels of student achievement because they have highly selective admissions policies and high demand for entrance. Controlling for student characteristics such schools may not have high levels of performance, however. Should we conclude that these schools are not performing well? To do so ignores the success of these organizations at an essential activity: obtaining inputs.

Similar considerations hold for the other factors usually considered fixed. Many organizations have some measure of control over their technology, the quality of staff, levels of environmental support, etc. Just as it does not appear meaningful to evaluate output performance holding input quality constant, so too with these other factors.

Once we acknowledge that all these factors are subject to organizational strategy and action, we are faced with a serious methodological problem. All of the variables that appear in the conventional analyses are endogenous, i.e., causally dependent on other variables in the model. For example, the quality of inputs may be functions of the expenditure on inputs which is a function of output performance. If none of these factors is causally prior or exogenous, it is extremely difficult to obtain unique estimates of any relevant causal effects in the system. In the technical language of econometrics, the system is underidentified. To remedy this situation one must make a considerable number of strong assumptions. Unfortunately the existing theories of organizational performance (and effectiveness) do not provide a basis for these assumptions.

The empirical literature on output performance does not measure up well when confronted with this list of methodological complications. To some extent the apparent lack of cumulation in the empirical literature on comparative effectiveness can no doubt be traced to a general failure to solve these problems. Each problem discussed above appears solvable. What is required is imaginative research design combined with careful model specification. We do not mean to suggest that the effort required is minor. It is not. It appears that it will take a considerable alteration in the way in which we approach comparative organizational research to solve the problems. But we believe that it can be done.

It is tempting to blame the lack of cumulative knowledge about the effects of organizational characteristics on effectiveness on the methodological problems just discussed and on difficulties in conceptualizing and measuring organizational goals. If this were so our message would be guardedly optimistic. But, while we agree that these issues contribute in important ways to the lack of cumulation in the comparative literature on effectiveness, we think that there are also more fundamental problems. We turn now to a discussion of the meaninglessness of comparisons of effectiveness (goal-attainment) across organizations.

C. Goal-Effectiveness

As noted earlier, (goal) effectiveness considerations involve comparisons of outcomes (e.g., output performance) with goals. Thus all of the conceptual and methodological problems discussed to this point also bear on the meaning and measurement of effectiveness. In our view, these are the least serious problems that arise in the comparative analysis of organizational (goal) effectiveness. In this section, we treat what we consider more fundamental problems.

We have argued that organizational goals are always multiple, are often unobserved, and are relatively non-specific, particularly with respect to time considerations. That is to say, that the preference functions we call goals do not have very strong properties. They will usually provide no more than a very rough ordering of outcomes in terms of preference. It helps to envision a concrete example. Suppose we observe two schools (or school districts) that each contain three students, and that the goal of each district is to maximize pupil academic achievement. Suppose further that achievement takes on only three values 0, 1, 2, denoting poor, moderate, and high levels, respectively. We denote outcomes to each school by a vector of the form (0,1,0), say, where the first element is the outcome to the first pupil, etc. Clearly, the schools prefer (2,2,2) to all other outcomes, and prefer any other outcome to (0,0,0). This minimal information enables us to make logically consistent comparisons of goal attainment across schools. If, for example, both schools begin at (1,1,1), and school A goes to (2,2,2) and B goes to (0,0,0), we conclude correctly that A has done a better job of goal attainment than B. These comparisons are permitted by the relatively weak goal statement because we observed change over time, and we considered only the extreme outcomes. Let us examine these two restrictions in more depth.

Suppose A has outcomes (2,2,2) and B has (1,1,1). Is A better off in terms of goal attainment than B? It is not obvious. We can draw such a conclusion only when the two preference functions are comparable and measurable ("cardinal" in the usual designation). The same sort of issue arises in assessments of change. Suppose A goes from (1,1,1) to (2,2,2) and B goes from (0,0,0) to (1,1,1). Are these equal gains in goal attainment (effectiveness)? Only if the utility functions are cardinal and inter-organizationally comparable.

The comparisons are often more complex than has been indicated so far. Suppose both schools begin at (1,1,1) and A shifts to (2,1,0) while B remains the same. Are the two schools equal in effectiveness (assuming, as we probably should not, that they were equal initially)? It depends on the nature of the goals or preference functions. One organization may be indifferent to such a change; it may focus only on average outcomes across pupils. Other schools may consider such a change an improvement in the sense that having one superior outcome is preferable to having none (regardless of what happens to the distribution); others may consider this new outcome ~~an~~ inferior ~~outcome~~ in the sense that having one failure is inferior to having none. The general point is that the public goals do not provide enough information to even order (in a strict order) all of the outcomes that fall between the extremes. We think this problem is general. For many organizations (even those with single goals), public goal statements often do not provide enough detail on the nature of the preferences to do more than indicate a weak ordering (a semi-order) of outcomes. If so, inter-organizational comparisons of effectiveness will be justified only for the most extreme changes.

Some of these difficulties arise because in the case of schools, it is meaningful to consider the outcomes to each pupil. Other organizations orient only to total or average outcomes rather than to the details of the distribution across subunits or objects worked on. For example, the classical firm orients only to total profit. In such cases, over-time comparisons of goal attainment are simplified in the sense that many of the intermediate cases collapse into single outcomes (if (2,1,0) and (1,1,1) denote the profits of subunits of a firm, the two outcomes are not differentiated). However, even in this simpler case, static comparisons of effectiveness are not justified. Nor can we conclude that a given change in profitability (say, a 5% gain) in two firms yields equal "utility" in each. To do this, we need much more information about the precise form of the preference or utility functions — we need cardinal preferences.

To this point, we have assumed single goals. Suppose the two schools also have as public goals maximizing the socialization of certain values. For each pupil, we must consider not only achievement but also values. That is, the set of outcomes of interest now consist of an achievement score and a values score for each pupil. Suppose that values are also measured on a three point scale so that the most preferred outcome is [(2,2),(2,2)(2,2)], and the least preferred outcome is [(0,0),(0,0),(0,0)]. We now have a very large number of intermediate cases. How, for example, do we compare [(2,0),(2,0),(2,0)] with [(0,2),(0,2)(0,2)]? To be able to do so requires that we know the precise weighting of each type of outcome (actually, we must know the precise weighting of each combination of outcomes). Again, *we must have cardinal preferences* in a situation in which the public goals give us no more than a very weak ordering of outcomes.

We believe this problem too is quite general. Above, we argued that all organizations (firms, hospitals, etc.) can be seen to have multiple goals in the sense that different units or levels each form their own distinct goals, and in the sense that timing considerations differ across organizational units and participants. For any concrete organization, it is a Herculean task to accumulate enough information to rank order all outcomes in terms of preferences of the organization as a corporate entity, all of these collectivities and units taken together (to do so also requires that we somehow weight the preferences of the corporate organization and of all the other units). It is inconceivable that we will obtain enough information to measure also the distances (in terms of preferences) among outcomes and a zero point (required for cardinality).

The problem as we have sketched it to this point is primarily an information problem. Many economists would argue that the problem of effectiveness comparisons is more fundamental. We are considering what amount to inter-personal comparisons of utility (where the persons may be literal persons, groups, or organizations as corporate entities). Although the matter is still the subject of some dispute, most micro-economists have long since concluded that such comparisons are not logically justified (see Arrow 1951 for a classic statement of this position). The problem has been rendered mute (except perhaps in welfare economics; see Rothenberg 1966) by the discovery that a satisfactory micro-economics (consumer theory, theory of the firm, etc.) can be constructed on the basis of ordinal utilities (ordinal indifference mappings). In fact, it is possible (see, for example, Henderson and Quandt 1971) because ~~that~~ modern consumer theory is unaffected by monotonic transformations of the preference functions. Therefore, there is no point in attempting to justify stronger utility arguments.

We believe that the field of organizations faces a similar situation. As we argued earlier, most organizational analysts need to employ goals (preferences) to relate environmental states to organizational action, in the same way that economists use preferences to relate price changes to consumer behavior. Comparative statements concerning goal attainment encounter all the difficulties that complicate inter-personal comparisons of utilities. In addition, goal functions for organizations are more complex. As a result of the shifting nature of goals within units of organizations, a set of additional complications arises in comparing goal attainment across organizations.

For the reasons just given, it is ~~usually~~ seldom possible ~~impossible~~ to compare inter-organizational utilities. As a consequence, students of effectiveness will tend to make such choices on the basis of their own personal values or those of the people who are paying them. Because there is no objective way to determine which goals are to receive emphasis in evaluating effectiveness, researchers will fall back on the arbitrary criterion of audience. So when evaluating a business firm for owners, profit will be the overwhelming criterion. When evaluating the same organizations from the members' point of view, member satisfactions will be emphasized.

Differences in goals also imply different operating functions to be specified in determining effectiveness. In most formulations, it is assumed that some overall or typical level of performance will be used in evaluation. It is clear, however, that this logic is appropriate for some kinds of organizations but not for others. Business firms, for instance, may be evaluated by total profits, or by earning per share. From this perspective, it makes no difference whether the organization made very high profits from one sale but lost on all others, or made a small but consistent rate of return on all its ventures. Often, however, organizations' one failure may be more important than a thousand successes. Custodial prisons, for example, are evaluated on the basis of the failures (Cloward et al. 1960). One escaped murderer costs more than a thousand successful incarcerations. Other organizations are evaluated in terms of infrequent but important successes. A research-oriented university which produces one Nobel Prize every five years will probably be judged effective, even if many of its researchers are doing pedestrian work. A competing university which does good solid but undistinguished research will probably be judged less effective if for no other reason than that those who do the judging are often unaware of good but undistinguished research. Finally, we can add that while some organizations are judged by the typical level of success, others are judged by the variance of success. Any organization which is judged in terms of justice will be evaluated in this way. So, as Weber pointed out (1947), the public goal of legal-rational bureaucracies is to treat each case the same. They operate according to standards of performance. Success is defined in terms of standardized outcomes, and variation is abhorred. The logic, based upon calculability of performance, stresses this invariant performance. Consequently, a failure is not counterbalanced by a success. A letter delivered with particular rapidity does not counter-balance a letter delivered with particular delay. Consequently, organizational characteristics which encourage effective performance for one organization may hinder it in another.

Many of the problems we have been discussing vanish when we drop the comparative focus. An investigator can certainly draw meaningful conclusions concerning effectiveness of certain elements of organizational design when interest is focused on the performance of a single organization, and when the analyst has access to precise information on the nature of organizational goals. This is the situation we earlier characterized as the engineering application. To repeat, we see no reason why social engineers ought not to couch their arguments in terms of effectiveness comparisons over time for single organizations.

For more comparative purposes oriented towards testing general and abstract propositions concerning organizational structure and behavior, we see no gain from utilizing the concept of effectiveness. Rather, we see an endless cycle of dispute over terminology and an inability to obtain inter-observer reliability across comparative studies of effectiveness. Similarly, it is almost impossible to prevent value statements creeping into effectiveness assessments in a way that impedes theoretical development and empirical cumulation.

III. Organizational Survival

When we began this discussion we distinguished between goal attainment and survival as ways of conceptualizing organizational effectiveness. In turning to the latter, we do not completely abandon the former. If we define the failure to survive as structural dissolution, it is not true that organizations which fail to achieve their goals or fail to make progress in that direction will fail to survive. Nor is it true that organizational "winners" which achieve their goals will survive.[4] In fact, being too successful may result in a loss of mission which results in survival failure - a thankless world. Corporate apologists sometimes mean that breaking up monopolists is "punishing success." Finally, we do not believe that we profit very much from assuming that loss of inputs results in failure to survive. Of course the absence of inputs will eventually have such a result. But most organizations which fall apart do so long before they reach this point. Some resources are more important than others and some are easily made up by conversion processes. Money for example is an extremely general resource which is translatable into many other scarce system inputs. We can say, however, that when there is a simultaneous decline in the various resources used by the organization, its existence is imperiled.

One set of circumstances under which this general decline is to be expected involves legitimacy crises. When organizational activities are normatively proscribed by the social system within which they take place, the consequences are likely to be many and varied. For example, the United States Army experienced such a catastrophe during the late 1960's when the Vietnam War received widespread criticism. It became difficult to obtain money for new weapons systems; attracting personnel became difficult; allies abroad began withholding cooperation. The consequences to be expected are very much the same as those which Stinchcombe (1965) described with the term "liability of newness." It is in this sense that the goals basis of definition retains its importance. When organizations are founded, a claim on resources is attempted through a statement of purpose. When the purposes are new or unique, when the method of pursuing them is different, or when alternative channels for achieving that purpose already exist, the claim will receive a rough reception.

The pretense of goal attainment is fundamental for acquiring legitimacy and other resources. Whether the organization in question actually pursues the goals is significant primarily as it serves as possible support for the pretense. In this sense the conceptualization of organizations as goal-directed devices is useful. Organizations are social systems whose creation is accomplished through a public affirmation of goal-seeking behavior. This announcement of intention serves as a legitimating device and as the basis of a claim on resources including participation. Public goal claims distinguish this form of social organization from others. The subject of this paper, effectiveness, is itself important as it relates to the pretense of goal-attainment. The goals themselves carry implications for assessment.

These remarks suggest that organizations with diffuse goals should be best able to protect themselves from adverse assessments. When the success of a given kind of organization is difficult to assess internally, it will also probably be difficult to assess from the outside. In particular, organizations with diffuse and/or multiple goals can always attempt to shift attention from some dimension on which performance appears substandard to some other dimension of assessment. In view of the advantages of having diffuse and multiple goals, why do any organizations have specific and limited goals? The answer is simple. It is extremely difficult to make diffuse goal claims stick. Only under rather exceptional circumstances (e.g. widely perceived societal crisis, absence of competition in some period, etc.) would we expect new organizations to be able to activate such diffuse claims for environmental resources. It seems much more likely that successful organizations with narrow goals may attempt to redefine their goals to make them more diffuse when and if they gain substantial control over their environments.

It is similarly difficult to establish a claim to very diverse and large amounts of resources. One of the dimensions on which organizations specialize or generalize in their resource-acquiring strategies is the claim on legitimacy. It is easier to persuade others that your intention is to perform certain narrowly defined tasks than it is to persuade others that you can and will perform every conceivable task. But making such a claim exposes the organization to greater risk. If one organization can make such a claim easily, so can others. And as is usually the case with specialism, the environment may change in such a way that any narrow statement of purpose may be only briefly persuasive.

When one focuses on the survival of individual organizations, the implicit assumption is that survival is a goal - that organizations try to survive. Unfortunately, this reintroduces many of the problems already discussed. Over time organizations that survive are all effective. This is because the survival criterion is binary. Levels of analysis are still problematic because passing the buck up and down the hierarchy is almost certain to obscure the issue of where causal processes are operating. Corporate decisions made at the highest level sometimes manifest themselves at low levels in ways unanticipated by the decision-makers. The system boundary problem is exacerbated because failing organizations are likely to form coalitions and other inter-organizational associations to provide for a mutual defense. The result is that the separate organizations lose their unit properties and the consequences of their actions are shifted upward toward the collectivity (Hannan and Freeman, 1976; Freeman, 1976). Finally, the endogeneity problem continues as failure to survive is often anticipated by other organizations which withhold resources. Bankers and suppliers refuse to extend credit as bankruptcy approaches. Conversely, a history of survival is often the best argument for resource acquisition.

Finally, we should also make it clear that using resource acquisition as a surrogate for survival requires a well-established theory of organizational death. That is, one must have some way of deciding how much of one resource makes up for a given loss of a quantity of another. One would need a theory which would describe the dynamics of the process.

Another way of using survival logic to study effectiveness would be by examining selection phenomena - survival in _a population_ of organizations. The primary benefit to be derived from this approach is that one does not have to attribute survival as a goal of individual organizations. Some may try harder to survive than others just as some members of the California Condor population may be trying harder than others. One does not have to introduce willfulness at all. Instead, one estimates the survival potential of the population by examining shifts in the distribution of organizational characteristics as organizations disappear and are replaced by new organizations.

This approach avoids the issue of whose goals the analyst selects for use in evaluation. For example, we can imagine two classes of organizations which are becoming increasingly scarce: Mom and Pop groceries and locally-owned banks. Mom and Pop may not be happy that the supermarkets and Seven-Eleven stores are squeezing them out. Their profits are low, their prices are high, the variety they offer is limited, and their establishemnt is poorly lit. Shoppers cast their dollar votes for the competition. Everyone concluded that their organizations were ineffective. The local banker, on the other hand, may succeed himself to death. That is, his bank may be growing yearly, returning a handsome profit to its owners, keeping its employees smiling, and serving its community well. However, it is its very profitability which leads a giant bank to make an offer that the owners cannot refuse. They sell out and make large capital gains. The new owner operates the local office impersonally and its emplyees satisfaction with working conditions declines. By "redlining" half the town, it engenders community hostility as assets are sent to other branch offices for investment. One would have a difficult time determining whether the new office is effective. Our point, however, is that these two situations are both examples of selection mechanisms which result in changes in the mix of organizations in the environment. Across large populations of organizations displaying those forms one could study the consequences of the characteristics which comprise them in terms of survival failure and this could generate abstract, logically consistent and falsifiable (i.e., scientific) theories. And they could help one design organizations more likely to survive.

New complications emerge from this approach as well, however, First and most obviously, it does not directly serve the interest of social engineers who want to decide how well some particular organization is doing. Such assistance will be rendered only if the scientific effort is successful. Second, it depends on the ability of the theorist to define and identify organizational forms that are empirically distinguishable.[5] Thirdly, it does not distinguish between the two types of organizational death referred to in our example: death by failure to compete effectively and death from successful competition (that leads to acquisition or perhaps even to a change in organizational form). In general the population perspective does not distinguish among the forms of environmental action that lead to systematic change in the mix of organizations in ways that correspond to the presumed or actual goals of the members or rulers of the organizations. We have argued throughout that to make such normative distinctions in theoretical work confuses science on the one hand and engineering and social criticism on the other.

Conclusions

Although our discussion has ranged over a number of issues, our main argument is quite simple. Most analysts equate effectiveness with goal-attainment. We have pointed to well known problems in conceptualizing and measuring organizational goals. These involve the non-specificity of most organizational goals with respect to the weighting of multiple goals and the weighting of short-run versus long-run payoffs. We have also discussed methodological problems that commonly arise in comparative studies of organizational performance. These less widely recognized problems include temporal considerations, establishing organizational boundaries, choosing the proper levels of analysis at which to measure outcomes, and the endogenous nature of so many variables typically treated as exogenous.

We believe there is a more fundamental problem. Public goal statements are not specific enough to justify comparisons across organizations in terms of goal-attainment. The problem is very reminiscent of the difficulties in making inter-personal comparisons of utilities. The problem is so serious that we argue that we drop any pretense to scientific analysis of organizational effectiveness.

This conclusion applies only to attempts at the constructing and testing of abstract theories of organizations. It does not follow that effectiveness considerations should not play a role in engineering or social criticism. We do insist on making clear the distinctions between these types of concerns.

Shifting attention from goal-attainment to survival (or to adaptability) simplifies the situation slightly. Nonetheless, to the extent that survival is treated as a goal, the central problems remain. These can be avoided by ignoring goals and considering selection in populations of organizations. New conceptual and methodological problems arise; but the fundamental objection to comparisons is eliminated.

FOOTNOTES

1. The theory is a straightforward application of Levin's (1968) fitness set theory, which is a particularly evocative form of niche theory.
2. Bidwell and Kassarda's response further clouds the issue by refusing to acknowledge the distinction between an event and a proposition. Our criticism concerned relations or causal processes at certain levels of analysis. In their reply, they translate our concerns into statements about events at various levels of aggregation. As soon as our concerns are stated (as originally) in terms of relationships, one sees that they have not addressed the criticism.
3. Under exceptional circumstances, it is possible that the bias arising from grouping observations may offset the bias due to the presence of uncontrolled individual properties (correlated with structural properties) in models used to estimate effects from individual data. In such a case, one actually gains from using the wrong level of analysis (Hannan 1976). None of the cases under consideration fits the conditions for aggregation gain.
4. We are ignoring for the moment all that we have said about how difficult it is to assess the degree of goal attainment.
5. All of these issues are explored in depth in Hannan and Freeman (1976) and Freeman (1976).

REFERENCES

Aldrich, Howard and Jeffrey Pfeffer
1976 "Environments of organizations," *Annual Review of Sociology* (forthcoming).

Arrow, Kenneth
1951 *Social Choice and Individual Values*. New York: Wiley.

Barnard, Chester I.
1938 *The Functions of the Executive*. Cambridge, Mass.: Harvard U. Press.

Baumol, William J.
1959 *Business Behavior, Value and Growth*. New York: Macmillan.

Bidwell, Charles E. and John D. Kassarda
1975 "School district organization and student achievement," *American Sociological Review* 40:55-70.

1976 "Reply to Hannan, Freeman, and Meyer, and Alexander and Griffin," *American Sociological Review* 41:152-160.

Blalock, Hubert M. Jr.
1964 *Causal Inferences in Non-Experimental Research*. Chapel Hill, N.C.: U. of North Carolina Press.

Blau, Peter M. and W. Richard Scott
1962 *Formal Organizations*. San Francisco: Chandler.

Cyert, Richard M. and James G. March
1963 *A Behavioral Theory of the Firm*. Englewood Cliffs, N.J.: Prentice-Hall.

Etzioni, Amitai
1965 "Two approaches to organizational analysis: a critique and a suggestion," *Administrative Science Quarterly* 5:257-278.

Freeman, John
1975 "The unit problem in organization's research," paper read a meetings of The American Sociological Association.

1976 "Defining organizational boundaries," unpublished manuscript.

Galbraith, John Kenneth
1967 *The New Industrial State*. Boston: Houghton Mifflin.

Hannan, Michael T.
1976 "Aggregation gain reconsidered," paper read at meetings of the American Educational Research Association.

Hannan, Michael T. and John Freeman
1974 "Environment and the structure of organizations," paper read at annual meetings of American Sociological Association.

1976 "The population ecology of organizations," *American Journal of Sociology* (forthcoming).

Hannan, Michael T., John Freeman and John W. Meyer
1976 "Specification of models of organizational effectiveness," *American Sociological Review* 41 (February):136-144.

Hannan, Michael T., Alice Young and Francois Nielsen
1975 "Specification bias analysis of the effects of grouping of observations in multiple regression models," presented at annual meetings of American Educational Research Association.

Hawley, Amos
1950 *Human Ecology*. New York: Ronald Press.

1968 "Human ecology," pp. 328-337 in *International Encyclopedia of the Social Services* (ed. David L. Sills). New York: Macmillan.

Henderson, James and Richard Quandt
1971 *Microeconomic Theory: A Mathematical Approach* (2nd ed.). New York: McGraw-Hill.

Hirsch, Paul M.
1975 "Organizational effectiveness and the institutional environment," *Administrative Service Quarterly* 20:327-344.

Katz, Daniel and Robert L. Kahn
1966 *The Social Psychology of Organizations*. New York: Wiley.

Lawrence, Paul and Jay Lorsch
1969 *Organization and Environment*. Homeworth, Ill.: Irwin.

Levins, Richard
1968 *Evolution in Changing Environments*. Princeton: Princeton U. Press.

March, James G. and Herbert Simon
1958 *Organizations*. New York: Wiley.

Michels, Robert
1949 *Political Parties*. Glencoe, Ill.: Free Press (originally published 1911).

Rothenberg, Jerome
1961 *The Measurement of Social Welfare*. Englewood Cliffs, N.J.: Prentice-Hall.

Scott, W. Richard, Sanford M. Dornbusch, C. J. Evashwick, L. Magnani and I. Sogatun
1972 "Task conceptions and work arrangements," Research Memorandum 97: Stanford Center for Research and Development in Teaching.

Simon, Herbert A.
1957 *Administrative Behavior* (2nd ed.). New York: Macmillan.

1962 "The architecture of complexity," *Proceedings of American Philosophical Society* 106:467-482.

Simon, Herbert A. (continued)
1964 "On the concept of organizational goal," Administrative Science Quarterly 9:1-22.

Steers, Richard M.
1975 "Problems in the measurement of organizational effectiveness," Administrative Science Quarterly 20:546-558.

Stinchcombe, Arthur
1965 "Social structure and organizations," pp. 153-193 in Handbook of Organizations (ed. James G. March). Chicago: Rand-McNally.

Williamson, Oliver E.
1964 The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm. Chicago: Markham.

Yuchtman, E. and Stanley E. Seashore
1967 "A system resource approach to organizational effectiveness," American Sociological Review 32:891-903.

The Effects of Correlated Units of Analysis: Choosing the Appropriate Unit*

Linda K. Glendening
National Institute of Education

Technical Report No. 12

September, 1976

Consortium on Methodology for Aggregating
Data in Educational Research

---

*This research was conducted under National Institute of Education Contract #NIE-C-74-0123

Vasquez Associates, Ltd.
1744 N. Farwell Ave.
Milwaukee, Wisconsin 53202

The Effects of Correlated Units of Analysis: Choosing the Appropriate Unit

Linda K. Glendening

National Institute of Education

September 1976

This paper is a revised version of the paper presented at the meeting of the American Educational Research Association held in San Francisco, April 1976.

STATEMENT OF THE PROBLEM

One of the assumptions underlying analysis of variance (ANOVA) is that units of analysis are independent of each other. Past educational research has, however, largely ignored this assumption. This neglect most probably is due to the fact that present definitions of independence are not only difficult to understand conceptually but also impossible to measure empirically. Regardless, it may very well be that violating the assumption of independence substantially affects the validity of any confidence statements based in part on that assumption and made regarding the hypothesized effects.

The intent of this study was to propose an operational definition of independence and, in light of that definition, study the problems with analyzing correlated units of analysis within the context of a two-level, hierarchically-nested ANOVA model, including only one outcome measure per subject. In addition, this study also considered problems with using a conditional testing procedure which includes a preliminary test of independence that selects the appropriate unit of analysis to use in testing the primary hypothesis of no treatment differences. Throughout this paper the effects of violating the assumption of independence will be discussed within the context of the general educational setting, using student as the disaggregate unit and classroom as the aggregate unit.

It seems that there are different reasons why individual students within a classroom may not respond to educational programs and tests independently of other students. Some classroom or grouping components that can affect the relationships between students' responses include:

- classroom environment effect
- teacher or instructional effect
- classmate effect

In order to see the effect of a common learning environment on the achievement of students, Steck (1966) randomly assigned thirty 7th graders to receive mathematical instruction in a group and another thirty to receive the instruction on a one-to-one basis. The presentation of the lesson was made by tape recorder to ensure similarity of treatment. Steck found that the variance of the scores of students who received the presentation as a group was significantly smaller than the variance of the scores of students who received the lesson individually. Steck's results suggest that in this particular study either common classroom experience decreased the variability of student responses and/or individualized instruction increased the variability of student responses. Clearly, however, this one study is not enough to conclude that the opposite condition (where classroom experience actually increases student variation and/or individualized instruction decreases student variation) does not occur.

Just how classroom or grouping components affect students may depend on such things as the student population. For example, the responses of kindergarteners may tend to be more related to each other than the responses of 12th graders. The type of program instruction might also make a difference. For example,a classroom discussion may tend to make students' responses more related to each other than a classroom lecture. Thus it can be seen that independence (or dependence) is not an all or nothing situation. Rather it is a matter of degree. At one extreme, observational units can be completely dependent upon each other. In this situation, N observational units are no more informative than one observational unit. At the other extreme, observational units can be completely independent (at least in theory). In this situation, N observational units give N pieces of non-overlapping information. Inbetween these two extremes is a complete continuum of dependency. It is this inbetween area that gives practicing statisticians headaches.

REVIEW OF LITERATURE

Definitions of Independence

Understanding the assumption of independence is prerequisite to studying the consequences of violating that assumption. This section contains a review of the varying definitions of independence contained in texts and papers dealing with statistical and research design issues. The conclusion is that, for the most part, these statistical sources have been too theoretical and mystical to inform practice. Below are examples of how different statisticians formally define independence. By themselves, these definitions seem inadequate as they are conflicting in the suggested causes of dependency and deficient in the assessment of dependency.

> The experimental errors must all be independent. That is, the probability that the error of any observation has a particular value must not depend on the values of the errors for other observations (Cochran, 1947).

> It is also assumed that the ϵij's are independent, both within each treatment level and across all treatment levels. If subjects are randomly assigned to treatment levels, the value of ϵij for any observation can be assumed independent of the values of ϵ_{ij} for other observations (Kirk, 1968, p. 52).

Before looking at additional definitions of independence, a distinction needs to be made between independence and random assignment. (Note that Kirk's definition equated independence with random assignment.) The condition of random assignment in a study is not synonymous with the condition of independence of observations. First, it should be mentioned that in theory independence can happen without random assignment. Students judgmentally assigned to any treatment condition can react to that treatment individually or independently of others assigned to that same treatment. Second, it can and does happen that units randomly assigned to treatment conditions do not, in fact, receive the treatments independently and thus the assigned units are not independent of each other. One example of this would be where children are randomly assigned to treatments but then all children assigned to any one treatment condition are treated as a group, e.g., they may all receive the treatment from the same teacher. Random assignment can only be counted upon to control disturbing or confounding variables that are present at the start of the study. Random assignment cannot control variables that are introduced, maybe only for convenience sake, into the experiment by the experimenter's manipulations, e.g., having one teacher instruct all students assigned to one treatment level.

A third definition of independence has been given by Draper and Smith (1966, p. 17). In discussing the general regression model, $Y_i = \beta_0 + \beta_1 X + \varepsilon_i$, for $i = 1,2,\ldots, n$, they make the following assumption:

> ε_i and ε_j are uncorrelated, $i \neq j$, so that cov $(\varepsilon_i, \varepsilon_j) = 0$. Thus ... Y_i and Y_j, $i \neq j$, are uncorrelated.

They then conclude that, given the above assumption and an additional assumption that ε_i is a normally distributed random variable, "ε_i, ε_j are not only uncorrelated but necessarily independent."

Draper and Smith's definition, in particular, seems especially hard to conceptualize. The difficulty lies in finding a way to measure the covariance between subjects when there is only one outcome score per subject. Within the single sample paradigm and under the condition that subjects are measured before selection, the cov $(\varepsilon_i, \varepsilon_j)$, for $i \neq j$, should always equal zero whenever subjects have been randomly selected from an infinitely large population. When subjects within a sample are more homogeneous on the dependent variable than subjects in the population, the cov $(\varepsilon_i, \varepsilon_j)$ will be greater than zero. The opposite condition of the cov $(\varepsilon_i, \varepsilon_j)$ being less than zero occurs whenever subjects within a sample are more heterogeneous on same outcome measure than subjects in the population.

Draper and Smith have suggested that one way observations would be independent of each other, before any treatment has taken place, is if the subjects are randomly selected from a normally distributed population. The condition of random selection from a normal population before treatment is not synonymous, however, with the condition of independence of observations after treatment. First, there is no reason that independence need be a function of the population being normally distributed. Theoretically, observations from a non-normal population can be independent of each other. Secondly, random selection does not insure independence of observations after a treatment has been effected. Random selection is concerned with the external validity of the experiment and cannot control variables that are later introduced into the study. For example, the situation where subjects randomly chosen were not, in fact, treated independently could disturb any pre-condition of independence that was due to subjects being randomly selected.

Cox (1958, p. 15) defines independence as the condition where "the observation on one unit is unaffected by the treatments applied to other units." He then goes on to say (1958, p. 19) that independence

> is the requirement that the observation on one unit should be unaffected by the particular assignment of treatments to the other units, i.e., that there is no "interference" between different units. In many experiments the different units are physically distinct and the assumption is automatically satisfied. If, however, the same object is used as a unit several times, or if different units are in physical contact, difficulties can arise.

Cox's definition of independence suggests that physical contact causes dependence. He states that if units are physically distinct the assumption of independence has been met. Within the educational setting, however, defining what "physically distinct" means is as difficult as defining what independence means. In looking over Cox's definition of independence, one must remember that Cox was mainly concerned with agricultural experiments. In these types of experiments, it happens to be much easier to control the interaction between different units. With educational studies, this is not usually the case. For example, the educational researcher cannot control what happens outside the school environment.

A fifth definition of independence concerns equating the experimental unit and the unit of analysis. If the statistical analysis of an experiment is to yield a valid confidence statement about the chances of drawing false conclusions from the data, the analytic unit should coincide with the experimental unit (Glass & Stanley, 1970; Peckham, Glass & Hopkins, 1969a, 1969b; Porter, 1972). This suggests that another definition of independence is that the condition of independence is satisfied when the unit of analysis is identical to the experimental unit. A unit of analysis refers to the smallest observational unit (or data point) which in the data analysis is to be considered distinct from other observational units.

Cox (1958, p. 2) has given the following definition of experimental unit:

> The formal definition of an experimental unit is that it corresponds to the smallest division of the experimental material such that any two units may receive different treatments in the actual experiment.

Cox then goes on to say that it is very "desirable" that experimental units also respond independently of one another. Peckham et al. have defined the experimental unit as:

> The experimental units are the smallest divisions of the collection of the experimental subjects which have been randomly assigned to the different conditions in the experiment and which have responded independently of each other for the duration of the experiment or which, if allowed to interact during the experimental period, have had the influence of all extraneous variables controlled through randomization.

Most of the above definitions of the condition of independence are not complete in that they are written only within the context of experimental situations. Definitions of independence should include correlational situations as well as experimental situations as the assumption of independence is made in correlational studies as well as in experimental studies. It is also interesting to note that some of the definitions of independence are procedural and imply causes (i.e., Cox and Kirk). Other definitions are given only in terms of outcomes or effects (i.e., Cochran and Draper and Smith).

Threats to Independence

Peckham et al. (1969a, 1969b) suggest two different ways that group influence can exert a dependency among units within a group on one or more dimensions. The first of these is an additive effect, which can raise or lower the group mean by tending to raise or lower the score of each unit within the group by a constant amount. This additive effect can influence the variability of classroom means, but at the same time do nothing to the variability of students' scores within the classroom. In other words, an additive type of dependency would disturb the predictability of the between class variance from the within class variance by affecting the former. The second type of group influence is what Peckham et al. call a proportional effect which can leave the mean performance of a group unchanged but can have a marked effect on the variability of responses within the group. This type of dependence disturbs the predictability of the between class variance from the within class variance, but this time affecting the latter.

These two general types of group influence defined by Peckham et al. can be broken into four conditions of dependence which systematically affect either the variation of students within classrooms or the variation of students between classrooms:

- An additive type of dependency which decreases the variation between classes,
- An additive type of dependency which increases the variation between classes,
- A proportional type of dependency which decreases the variation within classrooms and
- A proportional type of dependency which increases the variation within classrooms.

In real world situations it is possible that any one of the four dependency conditions suggested above might occur simply by nonrandom assignment of students to classrooms.

A decrease in the between class variation could occur in a study where judgmental assignment of teachers to intact classrooms has taken place. A situation where principals have assigned particularly effective teachers to difficult classes and average teachers to the more motivated classes could have the effect of decreasing the between classroom variation perhaps without changing the within classroom variation. On the other hand, an increase of the between class variation could be a function of classroom composition affecting teachers' attitudes toward teaching. Some teachers may become extremely ineffective when assigned to a particularly difficult class. This could have the effect of decreasing their students' achievement by a constant amount. Other teachers when assigned to a particularly motivated class may enjoy their working conditions so much that the achievement level of all their students within the class is increased by a constant amount, apart from the effect of the treatment being investigated. A situation occurring such as this could have the effect of increasing the between classroom variability without affecting the within classroom variability.

A decrease of variation within classes could be due, for example, to teachers working hard to increase the achievement level of disadvantaged students while at the same time ignoring the achievement potential of the brighter students. A situation such as this could decrease the within classroom variability without similarly affecting the between class variability. An increase of within class variation could be a function of teachers paying more attention to the more capable students in the class while ignoring the slower students. A situation such as this could increase the within classroom variance without changing the between classroom variance.

Selection of Analytic Units

In order that any definition of independence be operational, it must be able to help the researcher choose between alternative units of analysis and/or realize the consequences of a wrong choice of unit. In this section, past theoretical research will be reviewed that suggests how using a unit of analysis inappropriate to the research design can affect the results of a study. Empirical research will be presented that suggests how statistical power can differ given two distinct units of analysis, both being appropriate to the research design. The studies reviewed suggest that the test of the hypothesis of no treatment effects may be affected by the choice of unit of analysis. The magnitude of treatment effects (α_i's), however, should not depend upon the choice of the unit of analysis. Logical arguments will also be reviewed that discuss whether or not research questions dealing with educational programs should focus on the individual student. That is, are educational research questions involving the individual student functional given the present educational system?

Statistical Arguments for Unit-Selection

Scheffé (1959) has discussed the effects of violating the assumption of independence when observations are serially correlated. He considered the single group case where the random variables Y_i and Y_{i+1}, for $i = 1,2,\ldots, n-1$, had a serial correlation equal to ρ and all other pairs of observations had a serial correlation equal to zero. Scheffé found that the effect of serial correlation can be serious on inferences about means. As the serial correlation goes from 0 to -0.4, the test of the hypothesis becomes very conservative. As the serial correlation goes from 0 to +0.4, the test of the hypothesis becomes very liberal. Measurements that are either close in time or close in space can be serially correlated. Scheffé's results seem applicable in educational research when successive measurements have been taken on each experimental unit.

Cochran (1947), on the other hand, considered a simple group comparison case where every pair of observational units within a treatment level had a simple correlation of ρ. With correlated (ρ) units of analysis, the error of the treatment total $(e_1 + e_2 + \ldots + e_n)$ should have a variance equal to $n\sigma^2 + n(n-1)\rho\sigma^2$, rather than $n\sigma^2$ which would be the variance of the treatment total when errors are uncorrelated. Consequently, with correlated observations, the true variance of the treatment mean is $\{\sigma^2 + (n-1)\rho\sigma^2\}/n$. However the variance of the treatment mean is estimated by calculating the sum of squared deviations within each treatment level and then pooling across treatment levels. The variance of this treatment mean is equal to $\sigma^2(1-\rho)/n$. Therefore Cochran concludes that when observations are positively correlated, the true variance of the means is underestimated. When observations are negatively correlated, the true variance of the treatment means is overestimated. This suggests that Cochran's conclusions are consistent with Scheffé's in that when observations are positively correlated, the actual alpha level for the test of no treatment effect will be larger than the nominal alpha level, indicating that the test is too liberal; and when observations are negatively correlated, the actual alpha level for the test of the null hypothesis will be smaller than the nominal alpha level, indicating that the test is too conservative. Cochran handled the problem of dependent units of analysis within the same context as Draper and Smith's definition of independence. It is very hard to conceptualize what it means to have pairs of units of analysis correlated. For example, how is the correlations measured when there is only one outcome measure per student?

Probably the one most persuasive argument used by educational researchers in selecting the individual within a group, rather than the group itself, as their unit of analysis is the well-ingrained notion that studies with few observations tend to have little power for detecting treatment differences. Peckham et al. (1969a, 1969b) considered the hypothetical case of having 200 subjects randomly assigned to one of eight groups and groups randomly assigned to one of two treatment conditions. Using Kirk's (1968, p. 107) formula for estimating the noncentrality parameter, ϕ , under the condition of independence of individual observations, they computed power estimates using both the group as the unit and the individual as the unit (Table 1). Table 1 shows that, under these conditions, only for small treatment effects is the power using the group as the analysis unit much less than the power using the individual as the analysis unit. Under the conditions hypothesized by Peckham et al. no systematic differences occur between subjects across groups within treatment levels and therefore the expected mean square for groups within treatments equals the expected mean square for subjects within group/treatment combinations. Because these two expected mean squares are equal, the two F tests, one using group as the unit of analysis and the other using individual as the unit of analysis, have identical noncentrality parameters (ϕ) and thus the difference in the power of the two F tests is totally a function of the difference in degrees of freedom associated with each F test.

Table 1

Power Computations Using Groups and Individuals as Units of Analysis, Given Individuals are Independent

Analysis Unit	d.f. (error)	Power ($\alpha = .05$) Treatment Effect ($\mu_1 - \mu_2$ in sigma units) .25	.50	.75
Individuals	198	.42	.94	.991
Groups	6	.25	.82	.987

Logical Arguments for Unit Selection

Working within the realm of education, a question that needs to be asked is can individual student responses ever legitimately be assumed to be independent of each other? Or in other words, is student ever the appropriate unit of analysis? Within traditional education, the classroom is most often the functional treatment unit. That is, most often it occurs that all students within the same classroom receive the same basic instructional treatment. For example, it usually is the case that all children within classroom A learn math instructed by the same method. Rarely does it happen that within one classroom, math is taught to some students using the new method and to other students using the traditional method. Wiley (1970) has stated that

> If the object of evaluation is a typical classroom instructional program where the instruction is received simultaneously by all students in the class, then the appropriate vehicle (or sampling unit) is the class and not the individual pupil.

Peckham et al. (1969a, 1969b), Raths (1967) and Wiley have stated that if, however, treatments are presented in the form of individualized instructional techniques, such as programmed learning texts, dependency between students is unlikely to occur since the students should be working through the program on their own. Thus in programmed instructional experiments such as this, students would be the proper units of analysis. On the other hand, Haney (1974) has suggested that, even if students are given individualized instruction within the same classroom area, students' responses may not be independent of each other. Even though student A is working in his own carrel, he may find out at recess or after school that student B is moving along on his work a lot faster than he, student A, is moving. This could, for example, make student A hurry through his work too fast and thus student B would be having a negative effect on student A's learning.

Some people argue that using classroom means as units of analysis rather than individual student observations deprives them of stratifying on student variables of interest and hence the researcher loses all possibilities of finding aptitude/treatment interactions (ATI's). In other words, using classroom as the unit of analysis prevents researchers from investigating treatment by student characteristic interactions. This need not be the case. Repeated measures designs (Winer, 1962) do allow researchers to test for aptitude/treatment interactions while still using the classroom as the unit of analysis.

The definitions of independence and the research studies on the selection of appropriate analytic units discussed thus far are not sufficient to inform practice on the assumption of independence and consequences of violating that assumption. What is needed is an operational definition that is measurable. As long as the condition of independence is conceived to be an "ideal" property, undefinable operationally and hence unmeasurable directly, researchers will have trouble not only validating the assumption itself but also agreeing among themselves as to what the appropriate unit of analysis should be for specific research studies. An operational definition can be surmised by returning to what was said earlier about threats to the assumption of independence.

AN OPERATIONAL DEFINITION OF INDEPENDENCE

Peckham et al. (1969a, 1969b) suggest two general threats to independence. The first threat (an additive effect) affects the variation of the group or aggregate variable. The second threat (a proportional effect) affects the variation of individual units within the aggregate variable. And given random assignment of subjects to groups, these two variabilities are related. That is, given random assignment of subjects to groups, the variance of the group means will equal the variance of the population of subjects divided by the number of subjects per group. Translating this into an educational example, if students (S) are randomly assigned to classrooms (C), classrooms randomly nested within treatment levels (T) and students remain independent throughout the ongoing treatments, then the variance of the random sampling distribution of classroom means (σ^2_C) should equal the variance of the student population within classes and treatments ($\sigma^2_{S:CT}$) divided by the number of students per classroom (s), i.e., $\sigma^2_C = \sigma^2_{S:CT}/s$.

The two threats to independence and the general rule described above for predicting between group variation from within group variation provided the basis for the following proposed operational definition of independence. It is proposed that independence be operationally defined as the condition that σ^2_C equals $\sigma^2_{S:CT}/s$, or equivalently the condition that the expected mean square of the grouping variable, $E(MS_{C:T})$, equals the expected mean square of individual units within the grouping variable, $E(MS_{S:CT})$. More generally, this definition states that individuals or disaggregated units can be treated as independent units whenever the variance of aggregate units is predictable from the grouping size and the variance of the disaggregate units.

The intraclass correlation coefficient (ρI), which measures the extent to which observations within the same group tend to be homogeneous relative to observations across different groups (Kirk, 1968, p. 126-127), can also be used to define independence. Computationally the intraclass correlation coefficient equals:

$$\rho I = \frac{E(MS_{C:T}) - E(MS_{S:CT})}{E(MS_{C:T}) + (s-1)E(MS_{S:CT})} = \frac{s\sigma^2{}_C - \sigma^2{}_{S:CT}}{s\sigma^2{}_C + (s-1)\sigma^2{}_{S:CT}}$$

This formula indicates that whenever $\sigma^2{}_C$ equals $\sigma^2{}_{S:CT}/s$, or equivalently whenever the $E(MS_{C:T})$ equals the $E(MS_{S:CT})$, the ρI will equal zero. Thus an intraclass correlation coefficient equal to zero also operationally defines independence.

Given these three equivalent indicators of independence, operationally there are only two distinguishable types of dependence. Positive dependence is that condition where $\sigma^2{}_C$ is greater than $\sigma^2{}_{S:CT}/s$. Equivalently, positive dependence occurs whenever the ratio of the $E(MS_{C:T})$ over the $E(MS_{S:CT})$ is greater than one or whenever the ρI is greater than zero. Positive dependence is maximal when scores within each group are identical (i.e., when the $\sigma^2{}_{S:CT}$ equals zero) and the scores differ only from group to group (i.e., when the $\sigma^2{}_C$ is greater than zero). Negative dependence is that condition where $\sigma^2{}_C$ is less than $\sigma^2{}_{S:CT}/s$. Equivalently, negative dependence occurs whenever the ratio of the $E(MS_{C:T})$ over the $E(MS_{S:CT})$ is less than one or whenever the ρI is less than zero. Negative dependence is maximal when group scores are identical (i.e., when the σ^2_C equals zero) and when scores within each group differ (i.e., when the $\sigma^2_{S:CT}$ is greater than zero).

Each type of dependency (positive and negative) can result from any one of the following:

(a) an additive effect,

(b) a proportional effect and

(c) nonrandom assignment which will most probably result in either an additive or a proportional effect.

It happens that factors which would by themselves cause dependency can occur simultaneously in an experiment so as to counterbalance each other leaving the σ^2_C equal to $\sigma^2_{S:CT}/s$. For example, if interactions between students caused both the between class variance and the within class variance to decrease, positive dependency would not show up at least so long as the $E(MS_{C:T})$ remained equal to the $E(MS_{S:CT})$. These situations, however, really do not matter; as will be seen later the only thing that upsets the random sampling distribution of the F statistic of treatment effects when individual students are the analytic unit is the situation where σ^2_C does not equal $\sigma^2_{S:CT}/s$, given that the normality and homoscedasticity assumptions hold.

Again it should be mentioned that random assignment of students to classrooms and treatments is not synonymous with independence. However the condition of nonrandom assignment of students to classrooms almost certainly flags the condition of dependence between students. That is, under nonrandom assignment of students to classrooms, it seems unlikely that the operational definition of independence of students will hold.

Importantly the proposed operational definition of independence seems to be a useful one within the typical educational framework of hierarchical designs (i.e., designs which have students nested within classrooms, classrooms nested within schools, etc.). Further, this definition captures the usual definitions of independence at least insofar as variations from them affect the nature of the data. Finally, this definition has the advantage of being readily estimable.

ANALYTIC RESULTS

Armed with an operational definition of independence which is readily measurable, it is useful to reconsider the consequences of using correlated analytic units. As stated previously, failure to have independent units of analysis may bias parameter estimates, and this in turn may alter both the actual significance level and power of a test.

The Effects of Dependence

The effects of violating the assumption of independence will be studied within the context of a balanced, hierarchically-nested design, which has students nested within classrooms and classrooms nested within programs or treatment levels. Throughout the discussion below, the condition will be set that there is only one outcome measure for each student, that observations on students within classes and within treatments are distributed normally and have homoscedastic variances.

Data fitting the general educational model described above can be analyzed using one of two analysis of variance models. Model A (Table 2) was defined as:

$$Y_{ijk} = \mu + \alpha_i + b_{j(i)} + e_{ijk}$$

where α_i represents the effect of treatment i, $b_{j(i)}$ represents the effect of classroom j within treatment i and e_{ijk} represents the error of the kth student observation within the jth classroom and the ith treatment. Model B (Table 3) was defined as:

$$Y_{ik} = \mu + \alpha_i + e_{ik}$$

where α_i again represents the effect of treatment i and e_{ik} represents the error of the kth student observation within the ith treatment. Model B differs from Model A in that Model B contains no classroom component. Classroom is the analytic unit for Model A; while student is the analytic unit for Model B. Any hierarchically-nested data which fits Model A can also be analyzed using Model B.

The expected mean square tables for Model A and Model B, which were defined using the Millman and Glass (1967) rules of thumb, indicate that each model can test the hypothesis that there are no treatment differences. For Model A the test is $F = MS_T/MS_{C:T}$; while for Model B the test is $F = MS_T/MS_{S:T}$. Regardless of whether students are independent or not or whether there is a treatment effect or not, the $E(MS_T)$ for Model A equals the $E(MS_T)$ for Model B. Because the two F tests mentioned above have computationally identical numerators, problems caused by having correlated units of analysis only become evident in comparing the denominators of the two F statistics.

If students themselves are operationally independent on the dependent variable and there are no true treatment effects ($\sigma^2_T = 0$), then all expected mean squares in Models A and B estimate $\sigma^2_{S:CT}$. One operational definition of independence on the dependent variable discussed above equated the $E(MS_{C:T})$ to the $E(MS_{S:CT})$. Then given independence of students within classes the following holds for Model A:

$$E(MS_{S:CT}) = E(MS_{C:T})$$

$$E(MS_{S:CT}) = \sigma^2_{S:CT} + s\sigma^2_{C:T}$$

$$\therefore \sigma^2_{C:T} = 0$$

Table 2

Expected Mean Squares for Model A

$$Y_{ijk} = \mu + \alpha_i + b_{j(i)} + e_{ijk}$$

Sources of Variation	d.f.	E(MS)*
Treatment (T)	t-1	$\sigma^2_{S:CT} + s\sigma^2_{C:T} + sc\sigma^2_T$
Classroom:T (C:T)	(c-1)t	$\sigma^2_{S:CT} + s\sigma^2_{C:T}$
Student:CT (S:CT)	(s-1)ct	$\sigma^2_{S:CT}$
Total	sct-1	

$*\sigma^2_T = \Sigma\alpha^2_i/(t-1)$

Table 3

Expected Mean Squares for Model B

$$Y_{ik} = \mu + \alpha_i + e_{ik}$$

Sources of Variation	d.f.	E(MS)*
Treatment (T)	t-1	$\sigma^2_{S:T} + sc\sigma^2_T$
Student:T (S:T)	(sc-1)t	$\sigma^2_{S:T}$
Total	sct-1	

$*\sigma^2_T = \Sigma\alpha^2_i/(t-1)$

Thus the $\sigma^2_{C:T}$ component in the $E(MS_{C:T})$ formula can also be used to define independence. That is, whenever the $\sigma^2_{C:T}$ component equals zero, observations on students can be considered independent observations. That the $E(MS_{S:T})$ in Model B estimates $\sigma^2_{S:CT}$, when given independent student observations, can best be seen by looking at the sources of variation that combine to form the $SS_{S:T}$:

$$SS_{S:T} = SS_{S:CT} + SS_{C:T}$$

$$E(SS_{S:T}) = E(SS_{S:CT} + SS_{C:T})$$

$$E(SS_{S:T}) = E(SS_{S:CT}) + E(SS_{C:T})$$

$$(sc-1)tE(MS_{S:T}) = (s-1)ctE(MS_{S:CT}) + (c-1)tE(MS_{C:T})$$

$$E(MS_{S:T}) = \{sc\sigma^2_{S:CT} - \sigma^2_{S:CT} + sc\sigma^2_{C:T} - s\sigma^2_{C:T}\}/(sc-1)$$

$$E(MS_{S:T}) = \sigma^2_{S:CT} + \frac{(c-1)s\sigma^2_{C:T}}{(sc-1)}$$

and if $\sigma^2_{C:T} = 0$

$$E(MS_{S:T}) = \sigma^2_{S:CT}$$

Classroom as Unit

Independence. Whenever classroom observations are operationally independent of each other, there are no treatment effects and the assumptions of normality and homoscedasticity hold for classroom observations, the test statistic $F = MS_T/MS_{C:T}$ in Model A will have a central F distribution with $(t-1)$ and $(c-1)t$ degrees of freedom. This is true, in fact, regardless of what the distribution of students within classes looks like or regardless of whether or not student observations within classrooms are independent.

Given independence of student responses within treatments, varying the number of students per class and/or the number of classes per treatment should have no affect on the actual significance level of the F test $MS_T/MS_{C:T}$. On the other hand, it is predictable that increasing the number of students per class and/or increasing the number of classes per treatment should increase the noncentrality parameter, defined as $\lambda = 1 + sc\sigma^2_T/(\sigma^2_{S:CT} + s\sigma^2_{C:T})$, and thus increase the power of the test $F = MS_T/MS_{C:T}$. Given independence of students, i.e., $\sigma^2_{C:T} = 0$, one can predict, by looking at the noncentrality parameter, that increasing the number of classes per treatment or the number of students per class should identically inflate the $E(MS_T)$ and at the same time have no affect on the $E(MS_{C:T})$. However, increasing the number of classes per treatment should increase the power of the test $F = MS_T/MS_{C:T}$ to greater than that gotten by increasing the number of students per classroom by the same amount. This is due to the fact that increasing the number of students per class has no affect on the degrees of freedom error when classroom is being used as the unit of analysis. Whereas increasing the number of classes per treatment does increase the degrees of freedom error.

Positive Dependence. Positive dependence between student observations has been defined as the condition where the $E(MS_{C:T})$ is greater than the $E(MS_{S:CT})$, or concurrently the condition where the σ^2_C is greater than $\sigma^2_{S:CT}/s$. Positive dependence can occur either because a proportional type of dependency has decreased the within classroom variation or because an additive type of dependency has increased the between classroom variation. The positive dependency condition is identical to the case where the cov $(\varepsilon_i, \varepsilon_j)$, or similarly the intraclass correlation coefficient, would be greater than zero, under Draper and Smith's definition of independence. It is also identical to the instances where Cochran and Scheffé speak of responses of subjects being positively correlated.

When positive dependency occurs, under the condition of no treatment effects, $F = MS_T/MS_{C:T}$ has a central F distribution with $(t-1)$ and $(c-1)t$ degrees of freedom, given that classroom observations within treatment levels are independent, homoscedastic and distributed normally. This is true because the $\sigma^2_{C:T}$ component, which is greater than zero when student responses are positively correlated, affects the $E(MS_T)$ and the $E(MS_{C:T})$ in Model A similarly, given there are no treatment effects. Increasing the number of student responses within each class and/or increasing the number of classes per treatment should have no affect on the actual significance level of the test $F = MS_T/MS_{C:T}$. On the other hand, however, it is predictable, given positive student dependency, that increasing the number of students per class and/or the number of classes per treatment will increase the noncentrality parameter of $F = MS_T/MS_{C:T}$. In addition, increasing the number of classes per treatment increases the degrees of freedom error, which, in turn, should increase the power. Given positive dependence, s and c do not have identical effects on the noncentrality parameter.

The power of $F = MS_T/MS_{C:T}$ and the method of manipulating the degree (how big $\sigma^2_{C:T}$ is) of positive dependence are confounded. If degree of positive dependence is defined by increasing the $E(MS_{C:T})$, just such an increase will decrease power. However, if degree of positive dependence is defined by decreasing the $E(MS_{S:CT})$, degree of positive dependence will not affect the power of $F = MS_T/MS_{C:T}$. Degree of positive dependence is not confounded with significance level and will in no way affect the significance level of the F statistic using class as the unit of analysis.

Negative Dependence. Negative dependence between student responses has been defined as the condition where the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$, or concurrently the condition where the σ^2_C is less than $\sigma^2_{S:CT}/s$. Negative dependence can occur either because a proportional type of dependency has increased the within classroom variation or because an additive type of dependency has decreased the between classroom variation. Again, when dealing with real world data, these two conditions of dependency are indistinguishable and can be considered as one. This particular type of dependency is identical both to the case where the cov $(\varepsilon_i, \varepsilon_j)$, or similarly the intraclass correlation coefficient, is less than zero, under Draper and Smith's definition of independence, and to the instances where Cochran and Scheffé speak of subjects being negatively correlated.

As with positive student dependency, the amount of negative student dependency should have no affect on the distribution of the statistic $F = MS_T/MS_{C:T}$, given the condition of no treatment effects, and neither should increasing the number of students per class. This is because the test $F = MS_T/MS_{C:T}$ uses classrooms as the unit of analysis rather than students. In addition, given negative dependence, increasing the number of classes per treatment should also not affect the actual significance level of this test of no treatment effects as, under the central case, increasing the number of classes per treatment should have no affect on the parameters $E(MS_T)$ and $E(MS_{C:T})$.

How negative dependence affects the power of the test $F = MS_T/MS_{C:T}$ cannot be inferred from the expected mean square tables as defined for Model A (Table 2). This is due to the fact that the $\sigma^2_{C:T}$ term, which when greater than zero defines positive dependency, cannot be defined as equal to less than zero. Thus because of this term, $\sigma^2_{C:T}$, the noncentrality parameter, $\lambda = 1 + sc\sigma^2_T/(\sigma^2_{S:CT} + s\sigma^2_{C:T})$, is undefinable given negative dependence of student responses. As with positive dependence, when the test of the hypothesis is $F = MS_T/MS_{C:T}$, method of manipulating degree of negative dependence is confounded with affect on the power of $F = MS_T/MS_{C:T}$. The $E(MS_T)$ becomes larger as both the number of students per class increases and the number of classes per treatment increases. Thus an increase in s and/or c also increases the power of the above test. In addition, increasing the number of classes per treatment will increase the degrees of freedom error, which, in turn, will increase the power of $F = MS_T/MS_{C:T}$.

Student as Unit

The expected mean square formulas in Table 2 indicate that the $E(MS_{C:T})$ is the appropriate error term in testing for treatment effects. This is true whether or not student responses within classes are independent of each other. Therefore in order to check the validity of the test $F = MS_T/MS_{S:T}$, which has student as the unit of analysis, when students are not independent of each other, the $E(MS_{S:T})$ in Model B (Table 3) can be compared to the $E(MS_{C:T})$ in Model A (Table 2). If these two expected mean squares are not equal, then the $E(MS_{S:T})$ has to be biased in some way.

Independence. Given independence of student responses, $F = MS_T/MS_{S:T}$ would be an appropriate test of treatment effects for, as indicated earlier, the $MS_{S:T}$ and the $MS_{C:T}$ then estimate the same parameter, $\sigma^2_{S:CT}$. So, given independence of student responses within treatments, varying the number of students per class and/or the number of classes per treatment should have no affect on the actual significance level of the F test $MS_T/MS_{S:T}$. Whenever student responses are operationally independent, there are no treatment effects and the assumptions of normality and homoscedasticity hold for observations on students within treatment levels, the test statistic $F = MS_T/MS_{S:T}$ will have a central F distribution with (t-1) and (sc-1)t degrees of freedom. On the other hand, it is predictable, by looking at the expected mean square formula for the $E(MS_T)$ in Model B (Table 3), that increasing the number of students per class and/or increasing the number of classes per treatment should increase the power of the test $F = MS_T/MS_{S:T}$. Under the independence condition, the two F tests,

$F = MS_T/MS_{C:T}$ and $F = MS_T/MS_{S:T}$, should differ only in their power, with $F = MS_T/MS_{S:T}$ being the more powerful test as it has more degrees of freedom error.

Positive Dependence. The effect of positive student dependency on the dependent variable can be seen by comparing the formulas for the $E(MS_{C:T})$ and the $E(MS_{S:T})$.

$$E(MS_{C:T}) = \sigma^2_{S:CT} + s\sigma^2_{C:T}$$
$$E(MS_{S:T}) = \sigma^2_{S:CT} + \frac{(c-1)s\sigma^2_{C:T}}{(sc-1)}$$

Positive dependence deflates the $E(MS_{S:T})$ to less than the $E(MS_{C:T})$ as when the $E(MS_{C:T})$ is greater than the $E(MS_{S:CT})$, $\sigma^2_{C:T}$ will be greater than zero. However, $\sigma^2_{C:T}$ carries less weight in the $E(MS_{S:T})$ formula than it does in the $E(MS_{C:T})$ formula. Thus it can be seen that for all values of c and s, the $E(MS_{S:T})$ should be smaller than the $E(MS_{C:T})$ and therefore $F = MS_T/MS_{S:T}$ should be too liberal a test. When the $E(MS_{C:T})$ is greater than the $E(MS_{S:CT})$ and there is no treatment effect, $F = MS_T/MS_{S:T}$ should not be distributed as a central F. Rather $F = MS_T/MS_{S:T}$ should have an F distribution lying to the right of the distribution for the same F statistic under the condition of students being independent. That the positive dependency condition should result in too liberal a test statistic, when student is the unit of analysis, is consistent with what Scheffé (1959) and Cochran (1947) concluded when observations are positively correlated.

Given positive dependence, the degree of liberalness of the test $F = MS_T/MS_{S:T}$ is monotonically related to the degree of positive dependence. That is, as the $\sigma^2_{C:T}$ increases beyond zero, the discrepancy between the $E(MS_{C:T})$ and the $E(MS_{S:T})$ increases and thus the degree of liberalness increases. Given any one positive dependence level, this liberalness should be reduced as c increases and increased as s increases. Because of the general liberalness of the $F = MS_T/MS_{S:T}$ test, given positive dependence, the power of that test should be spuriously high.

Negative Dependence. The effect of negative dependency when student is used as the unit of analysis cannot be seen by comparing the usual formulas for the $E(MS_{S:T})$ and the $E(MS_{C:T})$. This is true because the Millman and Glass (1967) rules of thumb used to generate the expected mean squares for Model A no longer hold. What happens to the magnitude of the $E(MS_{S:T})$ when there is negative dependence, i.e., when the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$? The $E(MS_{S:T})$ is inflated to greater than the $E(MS_{C:T})$. The following relationships show why:

$$SS_{S:T} = SS_{S:CT} + SS_{C:T}$$

$$(sc-1)tE(MS_{S:T}) = (s-1)ctE(MS_{S:CT}) + (c-1)tE(MS_{C:T})$$

$$E(MS_{S:T}) = \frac{(s-1)c}{(sc-1)} E(MS_{S:CT}) + \frac{(c-1)}{(sc-1)} E(MS_{C:T})$$

Whenever the $E(MS_{S:CT})$ is at all larger than the $E(MS_{C:T})$, the $E(MS_{S:T})$ is larger than the $E(MS_{C:T})$, since for s greater than or equal to 2, $(s-1)c$ is greater than $(c-1)$. Therefore the test $F = MS_T/MS_{S:T}$ under this dependency condition should give too conservative a test. This indicates that whenever the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$ and there is no treatment effect, $F = MS_T/MS_{S:T}$ should not be distributed as a central F with $(t-1)$ and $(sc-1)t$ degrees of freedom. Rather $F = MS_T/MS_{S:T}$ should have an F distribution lying to the left of the distribution for the same F statistic under the situation where the $E(MS_{S:CT})$ is equal to the $E(MS_{C:T})$, or concurrently where the $E(MS_{S:T})$ equals the $E(MS_{C:T})$. That the negative dependency condition should result in a too conservative a test statistic, when the unit of analysis is the student, is consistent with what Scheffé (1959) and Cochran (1947) concluded when observations are negatively correlated.

Given negative dependency, the degree of conservativeness of the test $F = MS_T/MS_{S:T}$ is monotonically related to the degree of negative dependence. That is, as the degree of negative dependence increases (as the σ^2_C becomes smaller and smaller relative to the ratio $\sigma^2_{S:CT}/s$), the discrepancy between the $E(MS_{C:T})$ and the $E(MS_{S:T})$ increases and thus the degree of conservativeness should increase. As for positive dependence, given any one level of negative dependence, this conservativeness should be reduced as c is increased and increased as s is increased. The power of the test $F = MS_T/MS_{S:T}$ should be spuriously reduced as the conservativeness is increased. However both an increase in c and an increase in s inflates the $E(MS_T)$ and gives the error term of students within treatments more degrees of freedom. Thus both increasing c and s should increase the power of the test $F = MS_T/MS_{S:T}$. Whether $F = MS_T/MS_{S:T}$ will end up having more power than $F = MS_T/MS_{C:T}$ depends on how conservative $F = MS_T/MS_{S:T}$ is and how many additional degrees of freedom having students as the unit of analysis, rather than classrooms, gives.

In summary, failure to have independent units of analysis, as operationally defined in this study, does seem to bias the parameter estimates of the students within treatment error term, $E(MS_{S:T})$. And this bias, in turn, does influence the empirical alpha and power of the test $F = MS_T/MS_{S:T}$. The magnitude of this bias, given different degrees of both positive and negative dependence, was studied empirically in the original study, but will only be summarized in this paper.

The Preliminary Test

There are available at least two tests of treatment effects given a hierarchically-nested design with students nested within classrooms and classrooms nested within treatments. These two tests may conveniently be referred to as the "never pool" and the "always pool" procedures. The "never pool" test, $F = MS_T/MS_{C:T}$, uses classroom observations, which in this study are independent of each other, homoscedastic and normally distributed, as the units of analysis. Even though dependence of student responses does not affect the actual significance level of this test, always using the aggregate variable as the unit of analysis restricts the degrees of freedom error which, in turn, limits the power of this test of treatment effects. This few degrees of freedom problem motivates the need for another test or testing procedure with possibly greater power. The second and so-called "always pool" test, $F = MS_T/MS_{S:T}$, uses disaggregate or individual student observations, which within treatments are homoscedastic, normally distributed but not necessarily independent of each other, as the units of analysis. This particular analysis model has more degrees of freedom error, but, on the other hand, it has been shown analytically that any dependence between student responses adversely affects the significance level of the test of treatment effects. Thus it seems that there is no one "best" choice of unit of analysis for all hierarchical designs. Rather, the best testing procedure would entail using student as the analytic unit when student responses are independent of each other, and for all other circumstances using classroom as the analytic unit. This suggests a need for some sort of conditional testing procedure in which the unit of analysis to be used in the primary test of treatment effects is determined by a preliminary test of whether or not disaggregate units are operationally independent of each other.

When there is a question as to the validity of an assumption within an experiment, a preliminary test of significance can be used to support or reject the validity of that assumption. The procedure followed is to view the assumption as a hypothesis which is testable. If this hypothesis that the assumption is true is rejected, one takes action as if the assumption were false. On the other hand, if this hypothesis that the assumption is true is not rejected, one takes action as if the assumption were, in fact, true. This sort of assumption testing procedure is sometimes used by researchers in analysis of covariance models for testing the equality of regression slopes within groups. Another example of its use includes testing the equality of variances in analysis of variance models when groups are of different sizes.

If the researcher has some a priori notion that his individual observations are independent, Peckham et al. (1969a, 1969b) and Poynor (1974) recommend using this preliminary test of significance procedure to choose the unit of analysis for the primary test of treatment differences. The two-staged procedure takes the following form. The researcher begins with Model A (Table 2) and examines the null hypothesis that the variation between classes is no different than the variation within classes, H_o: $E(MS_{C:T})$ equals $E(MS_{S:CT})$. If the hypothesis of this preliminary test is not rejected, the researcher adopts Model B (Table 3) and pools the two mean squares, $MS_{C:T}$ and $MS_{S:CT}$. Using the pooled $MS_{C:T}$ and $MS_{S:CT}$ as the error term in a test of treatment differences is identical to using individual observations as the units of analysis. On the other hand, if the hypothesis of the preliminary test, H_o: $E(MS_{C:T})$ equals $E(MS_{S:CT})$, is rejected, the researcher retains Model A in testing for treatment differences and by doing so selects the classroom as the appropriate unit of analysis. This type of conditional testing procedure can be claimed a success if the actual alpha level of the primary or conditional test remains equal to the theoretical alpha and if the power of the conditional or final F test is greater than the power of the unconditional or $F = MS_T/MS_{C:T}$ test.

Actually in describing this conditional testing procedure for choosing an appropriate unit of analysis, Peckham et al. considered that this procedure would only detect dependencies of the additive type, where a constant is added to or subtracted from an entire class. Furthermore, they, as well as Poynor, considered only the possibility of the expected mean square between classes, $E(MS_{C:T})$, being larger than the expected mean square within classes, $E(MS_{S:CT})$. These two limiting considerations are needlessly restrictive. First, as previously indicated in this paper, the additive and proportional types of dependency which Peckham et al. define are in reality indistinguishable. And second, the preliminary F test can actually be rejected for one of two reasons. The $E(MS_{C:T})$ can be greater than the $E(MS_{S:CT})$, which will happen when the σ^2_C is greater than $\sigma^2_{S:CT}/s$. This signifies a positive dependency condition. On the other hand, theoretically the $E(MS_{C:T})$ can also be less than the $E(MS_{S:CT})$, which will happen when the σ^2_C is less than $\sigma^2_{S:CT}/s$. This signifies a negative dependency condition. Thus this preliminary F test should be a two-tailed test rather than the usual one-tailed F test which Peckham et al. and Poynor recommend.

Determining that individual observations rather than group observations should be the correct unit of analysis based on the initial test of independence is, in fact, a questionable analysis procedure. If the primary test of no treatment differences is based on the results of the preliminary test of independence, then the F test for no treatment effects is a conditional test and not a regular F test. This makes the test statistic have an unknown conditional distribution (Kirk, 1968). The conditional F test statistic need not be distributed either as the regular or "always pool" F statistic with (t-1) and (sc-1)t degrees of freedom, nor as the regular or "never pool" F statistic with (t-1) and (c-1)t degrees of freedom.

Paull (1950) has investigated the distributional properties of the conditional or so-called "sometimes pool" F statistic where the $MS_{C:T}$ and the $MS_{S:CT}$ form a pooled error term in testing for treatment effects only when the $E(MS_{C:T})$ is significantly greater than the $E(MS_{S:CT})$, which is the same limited condition that Peckham et al. and Poynor talked about. Paull found that under the currently proposed condition of operational independence, where the $E(MS_{C:T})$ equals the $E(MS_{S:CT})$, the preliminary F test, designed to choose the appropriate unit of analysis to use for the primary F test of treatment effects, is effective in making the power of the conditional test greater than the power of the "never pool" or $F = MS_T/MS_{C:T}$ test. But given the general condition of positive dependence, Paull found that the conditional test was more liberal and less powerful than the unconditional test $F = MS_T/MS_{C:T}$. It should be noted that Paull compared the powers of the "sometimes pool" and "never pool" tests at equal empirical alpha levels. That is, he did not confound power with the liberalness of the "sometimes pool" or conditional F test. As the ratio of the $E(MS_{C:T})$ to the $E(MS_{S:CT})$ increased from equal to one, Paull found that the observed alpha level of the "sometimes pool" or conditional test increased to a maximum and then decreased slowly to being equal to the nominal alpha level. This occurs because there is usually little power to find very small degrees of dependence, or very small differences between the $E(MS_{C:T})$ and the $E(MS_{S:CT})$. This means that given very small degrees of dependence the preliminary test will most often signal the individual, rather than the group, as the appropriate unit of analysis. And using the individual as the analytic unit will make the primary test of treatment effects too liberal a test. As the degree of positive dependence increases though, the researcher rejects the null hypothesis of operational independence more often. This dictates using classroom as the unit of analysis more often in the test for treatment differences. This, in turn, suggests that as the degree of positive dependence increases, the distributional properties of the "sometimes pool" or conditional F test become more and more similar to the distributional properties of the "never pool" or unconditional $F = MS_T/MS_{C:T}$. From this, one can conclude that for great amounts of dependency, the conditional test is just fine because it simply becomes an unconditional, "never pool" F test, $F = MS_T/MS_{C:T}$.

Paull also found that the number of classes per treatment and the number of students per class clearly affected the magnitude of the distributional differences between the "sometimes pool" and the "never pool" F test, given positive dependence. Under the condition of positive dependence, a large number of classes per treatment is desirable in two respects. First, as c increases the preliminary test becomes more powerful and correctly identifies classrooms as the appropriate unit more often. And second, when pooling of mean square error terms is prescribed, the pooled mean square $MS_{S:T}$ is weighted in favor of the valid and correct mean square error, $MS_{C:T}$. As the number of students per class, s, increases the preliminary test $F = MS_{C:T}/MS_{S:CT}$ again becomes more powerful and thus correctly signals the classroom as the proper unit of analysis more often. However, counterbalancing this positive effect of increasing s when given positive dependence is the fact that increasing s gives more weight to the wrong error term, $MS_{S:CT}$, which is smaller than the valid error term, $MS_{C:T}$. Thus the effect on the primary F test of increasing the number of individuals per group is due to a combination of two factors and most importantly depends on how much larger than the $E(MS_{S:CT})$ the $E(MS_{C:T})$ is. Lastly, Paull considered the effect of increasing the nominal alpha level of the preliminary F test and found that, given positive dependence, the magnitude of the undesirable property (liberalness) of the conditional F test was reduced somewhat with just such an increase. There was, however, a critical alpha level above which increasing the alpha level of the preliminary F test resulted in the conditional F test becoming more liberal. In Paull's example, this critical alpha value was very large, around 0.77.

Paull finally comes up with recommending the following rule as when and when not to pool the two mean square error terms $MS_{C:T}$ and $MS_{S:CT}$. The rule entails pooling the two mean square error terms only if their ratio is less than $2F_{50}$, where F_{50} is the 50 percent point of the F distribution with (c-1) and (s-1)ct degrees of freedom. Paull claims that this pooling decision rule is one which tends to "stabilize the disturbances" between the distributions of the two statistics $F = MS_T/MS_{C:T}$ and $F = MS_T/MS_{S:T}$ given "intermediate" conditions of positive dependence while still taking advantage of a considerable portion of the possible gain in power of pooling given very low levels of positive dependence. The present author, however, questions this "rule of thumb." Ideally the researcher wants most not to reject the null preliminary hypothesis of operational independence, as when this null hypothesis is not rejected the degrees of freedom error and mean squares between classes ($MS_{C:T}$) and within classes ($MS_{S:CT}$) can be pooled. The error the researcher needs to guard most against is a Type II error (β), or not rejecting the null hypothesis when it in fact is false. One way to decrease the probability of a Type II error is to increase the probability of a Type I error (α). However, doing as Paull recommends and taking twice the critical value given a large alpha of .50 has the same effect as selecting a small alpha level in the first place.

Again as mentioned above, within a two level hierarchically-nested design with individual units nested within groups and groups nested within treatments, Paull studied the distributional properties of the conditional F test given the usual one- and upper-tailed only preliminary F test. These distributional properties were studied only under the condition of positive dependence. It has been noted, however, that this preliminary test of independence can be rejected for one of two reasons (the occurrence of positive dependence and the occurrence of negative dependence) and thus instead should be a two-tailed F test. Or, if negative dependence is suspected, a one- and lower-tailed only preliminary F test would seem an appropriate possibility. On the other hand, there is no reason to believe that the presence and direction of the distributional effects found by Paull by changing the four parameters -- the number of individuals per group, the number of groups per treatment, the degree of dependence and the nominal alpha level of the preliminary F test -- should differ given negative dependence and/or a two- or lower-tailed preliminary F test. It is predictable, however, that given these other conditions, the magnitude of these effects should change. The empirical estimates of the magnitude of these distributional effects was investigated in the original study using simulated data, but will only be summarized in this paper.

SIMULATION PROCEDURES

The present investigation has addressed and discussed two specific questions. First, what is the effect of using correlated units of analysis? That is, what happens to parameter estimates and the probability of Type I and II errors when the assumption of independence is violated? And second, what is the effect of using a preliminary test of independence to choose the unit of analysis for the primary test of treatment effects? Thus far the two questions have been presented and discussed analytically. As yet though no attempt has been made to describe the actual magnitude of effects, whose presence and direction were predicted in the preceding analytic section, and which is the whole purpose of the simulation study. Thus the simulation study was done to demonstrate the size of the distributional effects for situations held to be common in educational settings.

The actual estimated magnitudes of effects will not be discussed in any detail in this paper. However, the procedures employed to empirically study the magnitude of distributional effects will be summarized.

Simulation Parameters

As stated previously, the general research design considered in the present study was a balanced, hierarchically-nested design. There were two levels to the nesting - individuals were nested within groups and groups were nested within treatments. The design assumed there to be one outcome measure per subject. Data such as this can be analyzed using one of two analysis of variance models, which were described in the previous section and presented in Tables 2 and 3.

The number of treatment groups, t, was held constant at two. Both the number of classes per treatment, c, and the number of students per class, s, were allowed to vary so that possible trends in the sampling distributions of the F statistic could be investigated, as these two parameters increased. Three values of classes per treatment (2, 5 and 10) were selected. The sample size of students ranged from five observations per classroom to 12 to 20. The noncentral case was created by adding 0.4 within class standard deviation units to each student observation in one treatment group. All sampling distributions were based on 1000 samples.

Two competing methods for defining and manipulating levels of positive and negative dependence among units were considered, both of which seem equally valid. One method uses the ratio of the $E(MS_{C:T})$ over the $E(MS_{S:CT})$. This method allows the intraclass correlation coefficient (ρI) to vary as the number of students per class varies (i.e., the ρI naturally gets smaller as the number of students per class increases). The intraclass correlation coefficient could also have been used to define and manipulate levels of dependence. This alternative method would entail keeping the ρI at a constant value for each level of dependence regardless of how many classes there were per treatment or students there were per class, but would force either the $E(MS_{C:T})$ or the $E(MS_{S:CT})$ to vary as the number of students per class varied. The first method described above for defining and manipulating degrees of dependence, altering the relationship between the $E(MS_{C:T})$ and the $E(MS_{S:CT})$ by varying the $E(MS_{C:T})$, was selected for several reasons. First, this method keeps the conceptual population of students the same. That is, randomly deleting or adding students to any classroom does not affect the within class variability. Second, previous research (Paull, 1950) has used the ratio of the expected mean squares between groups to the expected mean squares within groups in order to study the effects of preliminary tests for pooling mean squares.

Two general types of dependence were studied. Positive dependence was defined as occuring whenever the ratio of the $E(MS_{C:T})$ to the $E(MS_{S:CT})$ was greater than one (similarly the ρI was positive). Negative dependence was defined as occuring whenever the ratio of the $E(MS_{C:T})$ to the $E(MS_{S:CT})$ was less than one (similarly the ρI was negative). Within each type of dependence, two degrees or levels of dependence were studied. The degree of positive and negative dependence was varied in order to study probable trends of disturbance in sampling distributions of F tests of treatment effects, given different degrees of dependence. Along with the condition of independence the four defined conditions of dependence were:

- Independence

  $E(MS_{C:T})/E(MS_{S:CT}) = 1$

- Positive Dependence

  $E(MS_{C:T})/E(MS_{S:CT}) = 2$

  $E(MS_{C:T})/E(MS_{S:CT}) = 3$

- Negative Dependence

  $E(MS_{C:T})/E(MS_{S:CT}) = .50$

  $E(MS_{C:T})/E(MS_{S:CT}) = .33$

The choice of particular degrees of positive and negative dependence was somewhat arbitrary. However, an attempt was made to investigate degrees of dependence which typically occur within educational research studies. Smith (1974) suggested that, for elementary school children, classroom variance usually accounts for anywhere between 20 and 50 percent of the student variation within treatment levels for achievement measures such as reading and arithmetic. For studies conducted within a "tight" regional area, the classroom variation most likely would account for approximately 20 percent of the student variation; while for studies conducted nationwide the classroom variation most likely would account for about 50 percent of the student variation.

Haney (1974), working with data from 14 Philadelphia schools in the Follow Through study, found that students (ungrouped) had a variance of 137 on Metropolitan Achievement Test total mathematics scores. When Haney randomly grouped eight students (which was, on the average, the number of students per actual group), the variance between random groups was 25, while the variance between actual groups (or classrooms) was 52, which was twice the size of the variance using random groupings. Haney's data seem consistent with Smith's suggested limits in that when using the intraclass correlation coefficient to calculate the percentage of student variance attributable to classroom differences, the random groups accounted for approximately six percent of the student variation, while the actual classrooms accounted for approximately 28 percent. However, in calculating the variances, Haney did not take into account Follow Through, non-Follow Through differences. If he had, his intraclass correlation coefficients would likely have been somewhat smaller.

The actual degrees of positive dependence chosen reflect fairly well Haney's data, using sample sizes which seem in the range of common usage. It is also desirable that for different numbers of students, the intraclass correlation coefficient remain relatively small, as it is with small to intermediate degrees of dependence that effects of dependence and effects of the preliminary test of independence seem most nebulose. As Table 4 indicates, the intraclass correlation coefficients used in the present study were in the small to intermediate range.

Table 4

Theoretical Intraclass Correlation Coefficients
for Selected Numbers of Students and Degrees of Dependence

	$E(MS_{C:T})/E(MS_{S:CT})$				
	.33	.50	1	2	3
s = 5	-.153	-.111	.000	.167	.286
s = 12	-.059	-.044	.000	.077	.143
s = 20	-.034	-.026	.000	.048	.091

Data Generation Routine

The generation of unit normal variates involved two steps. First, the power residue method was used to generate twelve pseudo-random numbers from a uniform distribution. Second, the twelve numbers were summed and linearly rescaled, via the Central Limit Theorem, to provide N(0,1) variates.

Four basic steps were used to create the dependent variable Y_{ijk} with a known degree of dependence, where k indexes a student observation within classroom j and treatment i. First, s number of N(0,1) variates (Y'_{ijk}) were summed and averaged to get a classroom mean $\overline{Y}'_{ij.}$. Second, within each classroom, the classroom mean was subtracted from each individual observation $(Y'_{ijk} - \overline{Y}'_{ij.})$. Third, the class means were adjusted by the square root of the dependence level, $E(MS_{C:T})/E(MS_{S:CT})$. And finally, the adjusted means were added back to their respective set of deviations. Thus the dependent variable was calculated as

$$Y_{ijk} = Y'_{ijk} - \overline{Y}'_{ij.} + \{\overline{Y}'_{ij.} \sqrt{E(MS_{C:T})/E(MS_{S:CT})}\}$$

SUMMARY AND CONCLUSIONS

The main purpose of this study was to propose an operational definition of independence of analytic units, and to use that definition in an investigation of the effects of violating the assumption of independence. A secondary purpose was to expand upon a conditional testing procedure that has been proposed in past research and to test its validity. This conditional testing procedure included a preliminary test of independence that was used to select the unit of analysis to use in testing the primary hypothesis of no treatment differences. How the number of individual units per group, the number of groups per treatment level and the type and degree of dependence affected both the validity of using correlated units of analysis and the consequences of using a conditional testing procedure were studied analytically. In addition, the size of the effects were estimated empirically. This empirical estimation was done to see if effects on alphas and powers were large enough to affect practice. In some respects, the parameter values specified for the Monte Carlo part of this study limit generalizations of empirical effects. However, care was taken to select parameter values held to be common in educational data.

It was proposed that independence of analytic units be operationally defined as: Disaggregate or ungrouped units can be considered as independent units whenever the variance of the aggregate or grouped units can be predicted from the grouping size and variance of the disaggregate units. This definition of independence can be translated into testing the equality of expected mean squares between and within groups. That is, given the above definition of independence, the expected mean square between groups, $E(MS_{C:T})$, should equal the expected mean square within groups, $E(MS_{S:CT})$, for the hierarchically-nested design considered in the study. Also under independence of responses between and within groups the intraclass correlation coefficient will equal zero. These congruent definitions of independence appear to be useful within the framework of any hierarchically-nested design having only one observation per subject and can be used to check the equality of any pair of levels given n hierarchically-nested dimensions.

Given the above operational definition of independence, two types of dependency are possible - positive and negative. Positive dependence was defined by the expected mean square between groups being larger than the expected mean square within groups, or similarly by the intraclass correlation coefficient being greater than zero. Negative dependence was defined by the expected mean square within groups being larger than the expected mean square between groups, or similarly by the intraclass correlation coefficient being less than zero. Both types of dependence can be caused by an additive effect (which influences the variation between groups), a proportional effect (which influences the variation within groups) and nonrandom assignment. Random assignment of students to classrooms does not, however, perforce erase the possibility of positive and/or negative dependence occurring.

Given independence of disaggregate units, either student or classroom can be used as the unit of analysis. Both the "never pool" test, $F = MS_T/MS_{C:T}$, and the "always pool" test, $F = MS_T/MS_{S:T}$, are appropriate tests of treatment effects. All empirical alphas for both tests were within 1.96 standard errors of their nominal values. $F = MS_T/MS_{S:T}$ is the preferable test, however, as it always had more power than did the test $F = MS_T/MS_{C:T}$. Given independence, the discrepancy in power between these two tests comes solely from the difference in degrees of freedom error for each of the F tests. Over all simulated conditions, the power of the "always pool" test exceeded that of the "never pool" test by an average of .081.

Given positive dependence, where the $E(MS_{C:T})$ is greater than the $E(MS_{S:CT})$, the proper unit of analysis is the grouped unit, classrooms. Using the disaggregate unit as the unit of analysis caused the pooled error term, $E(MS_{S:T})$, to be biased on the small side. And this attenuation caused the "always pool" test, $F = MS_T/MS_{S:T}$, to be too liberal. For both simulated degrees of positive dependence, none of the empirical alphas generated were within 1.96 standard errors of their respective nominal values. This suggests that the effect of positively correlated units can result in spurious significance that is of practical importance. The liberalness increased as the number of students per class increased and as the degree of positive dependence increased, i.e., the ratio of the $E(MS_{C:T})$ over the $E(MS_{S:CT})$ increased above one. The liberalness decreased as the number of classes per treatment increased. Because of the general liberalness of the "always pool" test, given positive dependence, the empirical power of that test was spuriously high.

Given negative dependence, where the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$, the correct unit of analysis is once again the grouped unit. Using the ungrouped unit as the unit of analysis caused the pooled error term, $E(MS_{S:T})$, to be biased on the high side. That is, the "always pool" test, $F = MS_T/MS_{S:T}$, was too conservative a test. For both simulated degrees of negative dependence, none of the empirical alphas were within 1.96 standard errors of their nominal alphas. This suggests that the effect of negatively correlated units of analysis can result in spurious lack of significance that is of practical importance. While the analytic analysis suggested that increasing the number of students per class should increase the conservativeness and increasing the number of classes per treatment should decrease any conservativeness, the empirical analysis found no clear trends. Decreasing the ratio of the $E(MS_{C:T})$ over the $E(MS_{S:CT})$ to below one (which is the same thing as increasing negative dependence within the data) did, however, increase the conservativeness of the "always pool" F statistic, as expected. The conservativeness found in the negatively dependent situations spuriously reduced the empirical power of the "always pool" test to such an extent that, in all simulated cases but one, the advantage of using the "always pool" test in the first place (more degrees of freedom error) was cancelled out.

What do the above analytic and empirical results suggest for the practitioner? They suggest that when dealing with educational data, in almost all cases, the grouped unit, such as classrooms, should be the unit of analysis. If, however, the data do happen to be independent of each other, it is clearly advantageous to use the individual unit as the unit of analysis. The results indicated quite convincingly that the F test is not robust to violations of the assumption of independence. This conclusion should be kept in mind both when designing and analyzing experiments as well as when interpreting the results from studies which have not followed the advice from this investigation. Closely tied to this, of course, is a real need to empirically determine the types and degree of dependence most common in real world, educational data.

One might next ask, Is it feasible to first do an initial test of independence, and then on the basis of that test choose a unit of analysis for the primary test of treatment effects? This study showed clearly, both analytically and empirically, that the answer to this question is no. Two criteria were used to judge the adequacy of the conditional testing procedure. First, the empirical alpha should be close to its respective nominal value. And second, the power of the conditional test should be greater than the power of the always correct, "never pool" test, $F = MS_T/MS_{C:T}$.

Only when independence of student responses occurred did the conditional testing procedure turn out to very effective. For that condition and given the two-tailed (H_0: $E(MS_{C:T}) = E(MS_{S:CT})$), the upper-tailed only (H_0: $E(MS_{C:T}) \leq E(MS_{S:CT})$) and the lower-tailed only (H_0: $E(MS_{C:T}) \geq E(MS_{S:CT})$) preliminary tests, 80% of the empirical conditional test alphas were within 1.96 standard errors of the nominal alphas and 96.5% of the powers of the conditional test were greater than comparable powers of the "never pool" test, $F = MS_T/MS_{C:T}$. Across the three types of preliminary tests and across all simulated conditions, the difference between the power of the "sometimes pool" F test and the "never pool" F test averaged .073, favoring the "sometimes pool" test. Given any of the three preliminary F tests, the discrepancy between the powers of the "sometimes pool" test and the powers of the "never pool" test decreased as the number of groups per treatment increased and as the alpha levels of the "sometimes pool" and "never pool" tests increased. On the other hand, the discrepancy increased as the number of students per group increased. For both the two-tailed and upper-tailed preliminary test situations, the discrepancy between the powers of the "sometimes pool" and the "never pool" tests decreased as the alpha level of the preliminary test increased. However, this trend (which was expected) was reversed when the lower-tailed preliminary test was used, most likely because the observed alpha levels tended to become too liberal as the preliminary test alpha level increased.

Given positive dependence, the conditional F test generally turned out to be too liberal a test and thus had spuriously high empirical power. Given either the two-tailed or upper-tailed preliminary tests, the liberalness of the "sometimes pool" statistic decreased as the number of groups per treatment increased, as the degree of positive dependence increased, as the alpha level of the preliminary test increased and as the alpha level of the "sometimes pool" test increased. Increasing the number of students per group had no simple effect on the liberalness of the "sometimes pool" test. Across the simulated designs considered and given the two-tailed preliminary F test of independence, 22% and 35.2% of the "sometimes pool" or conditional F tests were, however, robust to the occurrence of positive dependence where $E(MS_{C:T})/E(MS_{S:CT})$ equalled 2 and 3 respectively. For those simulated circumstances, where the "sometimes pool" test did empirically appear to be robust to the occurrence of positively correlated analytic units, the difference between the empirical powers of the "sometimes pool" and the "never pool" tests averaged only .014, favoring the "sometimes pool" test. Given positive dependence, the magnitude of effects for the conditional test, following either a two-tailed preliminary test done at 2α or an upper-tailed preliminary test done at α, were comparable. On the other hand, the "sometimes pool" F test statistic, given positive dependence and a lower-tailed preliminary test, was distributed as the $F = MS_T/MS_{S:T}$ test statistic as varying the alpha level of the preliminary test had negligible affect on the "sometimes pool" test statistic.

Given negative dependence, the conditional test was somewhat conservative and generally had less power than the "never pool" $F = MS_T/MS_{C:T}$ test. Given either the two-tailed or lower-tailed preliminary tests, the conservativeness of the "sometimes pool" statistic decreased as the degree of negative dependence increased, as the alpha level of the preliminary test increased and as the alpha level of the "sometimes pool" test decreased. And also as expected, increasing the number of students per group again had no simple effect on the conservativeness of the "sometimes pool" test. The analytic analysis suggested that increasing the number of groups per treatment should decrease the conservativeness of the "sometimes pool" test statistic, but the simulated data were weak in confirming this expected trend. Across the simulated designs considered and given the two-tailed preliminary F test, 60.8% and 72% of the "sometimes pool" or conditional F tests were empirically robust to the occurrence of negative dependence where $E(MS_{C:T})/E(MS_{S:CT})$ equalled .50 and .33 respectively. For those specific empirically robust instances, the difference between the empirical powers of the "sometimes pool" and the "never pool" tests averaged - .010, favoring the "never pool" test. Given the $E(MS_{C:T})$ was less than the $E(MS_{S:CT})$, the magnitude of effects for the conditional test, following either a two-tailed preliminary test done at 2α or a lower-tailed preliminary test done at α, were comparable. The "sometimes pool" F test statistic, given negative dependence and an upper-tailed preliminary test of independence, was distributed like the "always pool" F test statistic as varying the alpha level of the preliminary test had essentially no affect on the "sometimes pool" test statistic.

In summary, this study has shown that, if individual units are independent, the "always pool" $F = MS_T/MS_{S:T}$ test is "best". If individual data are not independent, the "never pool" $F = MS_T/MS_{C:T}$ test is "best". However, the problem is the researcher is never actually in the position of knowing before the analysis stage whether or not responses of subjects nested within groups are, in fact, independent responses. Keeping in mind the assumptions underlying this study, such as, classrooms are independent of each other (and in reality, one can question the validity of this assumption, but they should always be more independent of each other than students), this study has shown that across all circumstances it is best to use the classroom as the unit of analysis. And as a second choice, it is better to let a preliminary test select a unit of analysis than to always choose student as the unit. In so doing, one should always use a relatively large alpha level for the preliminary test.

REFERENCES

Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics, 1947, 3, 22-38.

Cox, D. R. Planning of experiments. New York: Wiley & Sons, 1958.

Draper, N. R., & Smith, H. Applied regression analysis. New York: Wiley & Sons, 1966.

Glass, G. V., & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall, 1970.

Haney, W. Units of analysis issues in the evaluation of Project Follow Through or There must be heresy in this some place (Contract No. OEC-0-74-0394). Cambridge, Mass.: Huron Institute, September 1974.

Kirk, R. E. Experimental design: Procedure for the behavioral sciences. Belmont, Calif.: Brooks/Cole, 1968.

Millman, J., & Glass, G. V. Rules of thumb for writing the ANOVA table. Journal of Educational Measurement, 1967, 4, 41-51.

Paull, A. E. On a preliminary test for pooling mean squares in the analysis of variance. Annals of Mathematical Statistics, 1950, 21, 539-556.

Peckham, P. D., Glass, G. V., & Hopkins, K. D. The experimental unit in statistical analysis. The Journal of Special Education, 1969, 3, 337-349. (a)

Peckham, P. D., Glass, G. V., & Hopkins, K. D. The experimental unit in statistical analysis: Comparative experiments with intact groups (No. 28). Boulder, Colo.: University of Colorado, Laboratory of Educational Research, March 1969. (b)

Poynor, H. Selecting units of analysis. In G. Borich (Ed.), Evaluating educational programs and products. Englewood Cliffs, NJ: Educational Technology Press, 1974.

Raths, J. The appropriate experimental unit. Educational Leadership, 1967; 263-266.

Scheffé, H. The analysis of variance. New York: Wiley & Sons, 1959.

Smith, M. Personal communication, October 21, 1974.

Steck, J. C. The independence of observations obtained in classroom research. Unpublished M. A. Thesis. College Park, Md.: University of Maryland, 1966.

Wiley, D. E. Design and analysis of evaluation studies. In M. C. Wittrock & D. E. Wiley (Ed.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart & Winston, 1970.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

The Effects of Education As An Institution*

John W. Meyer
Stanford University

Technical Report No. 13

August, 1976

Consortium on Methodology for Aggregating
Data in Educational Research

---

*This research was conducted under National Institute of Education Contract #NIE-C-74-0123

THE EFFECTS OF EDUCATION AS AN INSTITUTION*

John W. Meyer
Department of Sociology
Stanford University

August 1976

---

*This paper was prepared with funds from the National Institute of Education (Contract NIE-C-74-0123 to Vasquez Associates, Ltd.). The views expressed here are those of the author, not of the N.I.E. Some ideas developed here are presented in more limited form in Meyer (1972) and Meyer and Rubinson (1975). I am indebted to the advice and help of many colleagues: among them William Bowers, Christopher Chase-Dunn, Michael Hannan, David Kamens, Francisco Ramirez and Richard Rubinson.

THE EFFECTS OF EDUCATION AS AN INSTITUTION

Research on the effects of educational systems on society has for the most part been restricted to (1) studies of the effects of schools on students and (2) studies of the effects of education on the subsequent careers of students as individuals. Such studies have produced many findings of interest, but their focus on the individual effects of education has tended to obscure many social processes of interest from consideration. In this paper, we consider social effects of education which transcend the processes of individual socialization. We discuss the effects of education as an institutionalized system of allocation on both those who attend and those who do not attend the schools. In the course of this discussion, some interesting indirect socializing effects of education appear, and we arrive at an explanation of a major paradox of contemporary research on school effects--the finding that schooling makes enormous differences in individual outcomes, while particular schools make very little difference.

We then go on to consider the effects of education as an institution on society: The ways institutionalized educational systems may legitimate and transform other social institutions (apart from socializing and allocating processes). That is, we consider the effects of education, at the aggregate level, on the orientation and behavior of individuals and elites regardless of their own education.

In these arguments, we move away from the contemporary view of educational organization as a production system constructing elaborated individuals as products. We conceive of modern education as institutionalized rites transforming individual roles through powerful and elaborated initiation ceremonies, which reorganize society by building into it classes of formally reconstituted authorized personnel, understood to possess expanded and reconstituted bodies of authoritative knowledge.

THE TRADITIONAL SOCIALIZATION MODEL

Prevailing research on school effects is organized around a simple image of socialization in society: Schools install knowledge and skills, attitudes and values, in their students. These students then have a revised and expanded set of personal qualities enabling them to demand more from, and achieve more in, the role structure of modern society. As the competence and orientation of the personnel of society is expanded and modernized, so society as a larger system is modernized and expanded.

Three general propositions are at issue here, and make up a simple model which is diagrammed in Figure 1:

> Proposition 1 (Socialization): Schooled persons are socialized to expanded levels of knowledge and competence, and expanded levels of modern values or orientations.
>
> Proposition 2 (Socialization and Adult Competence): Early socialization to higher levels of knowledge, competence, and modern values or orientations creates higher levels of adult status and competence.
>
> Proposition 3 (Individual Competence and Social Progress): The expansion of the number of skilled adults expands the complexity and wealth of society and social institutions.

Research on Proposition 1 is rather clear cut. Children and youth in schools learn a good deal more, and acquire more expanded social capacities than those not in school even when background factors are controlled (see e.g., Holsinger, 1975; Plant, 1965). The main problem in the research on this subject is the finding that the particular school success and effectiveness.

ALLOCATION THEORIES

In reaction to the empirical weaknesses of traditional socialization theory, the view arises that apart from anything they have learned, people are allocated to adult roles involving authority and efficiency on the basis of education, in modern societies. Explanations of why this happens so consistently are vague--we discuss this issue later. The power of the argument arises from its obvious empirical validity. We all know that status positions in modern society are assigned on the basis of education.

Sometimes, as with civil service and professional positions (e.g., medicine, law, teaching), this is a matter of law. To teach in a high school, one must have an educational credential. Whether one knows anything or not is less relevant.

Often, credentialling rules are simply part of established organizational practice, as in the assignment of college and business school graduates to managerial positions and of others to working class jobs. Sometimes the whole process is informal, as in the inclination of juries and informal friendship groups to attend to the advice of their more educated members.

In any event, the relationship between education and social position--over and above socialization or learning--is quite direct. The line of research pursued by Blau and Duncan (1967) and Duncan, Featherman and Duncan (1972) shows large direct effects of education on status attainment, sometimes with ability measures held constant. It is clear that education plays a direct causal role in occupational transition even late into the individual's career (Blau and Duncan, 1967, Ch. 5)--decades after any direct socialization effects must have decayed or become outmoded.

Similarly, Almond and Verba (1963) and Inkeles and Smith (1974) show, with data on a number of countries, that education is uniformly the background factor most associated with political information, attitudes, and participation. These are, possibly, traditional socialization effects, but in the face of the evidence on the continuity over time of personal qualities, it seems easier to believe that direct educational allocation is involved.

The basic idea is clear:

> Proposition 4 (Educational Allocation): In modern societies, adult competence, success, and orientation are assigned to persons on the basis of education, holding constant what they may have learned in school.

The power of this idea lies in its implications for explaining previous findings and proposing new lines of research. Two major propositions follow immediately, if we simply assume that persons adapt to their actual and expected experience:

> Proposition 5 (Chartering): Students tend to adopt personal and social qualities which are appropriate to the social positions to which their schools are empowered to assign them.
>
> Proposition 6 (Lagged Socialization): Adults tend to adopt qualities appropriate to the roles to which their educational statuses have assigned them.

Proposition 5 offers a direct explanation of the most puzzling general research paradox in the sociology of American education: The finding that students with given backgrounds tend to learn a great deal in school when compared with non-students, but that variation in learning and socialization among schools--even though these schools differ greatly in structure and resources--are very small. The idea that schools serve as ritually chartered organizations (Meyer, 1970; Kamens, 1971, 1974) rather than production organizations produces the argument that schools with similar credentialling positions in society will produce similar effects, with little regard to variations in their structures. Since for many personnel assignment purposes all American high schools have similar status rights as do all colleges, variation in their effects should be slight. But because all high schools are chartered to create "high school graduates"--a critical status in our society--all of them tend to produce marked effects on students (in contrast to non-students).

Thus, the educational contexts which substantially affect change and learning in students do not usually include the school. They include contexts which are distinctly chartered:

-- Schooling per se: Life prospects (and hence student changes) are vitally affected by being in an institution chartered as a school.

-- Type of school, when these types are differently chartered: Himmelweit and Swift (1969) and Kerckhoff (1975) show marked differences in outcomes for similar British students between grammar and secondary modern schools. American researchers have not looked for differences in expectations between initially similar students in general and vocational high schools. Some studies show distinct occupational effects of teachers colleges and engineering schools (Astin & Panos, 1969)

-- Curriculum, when these are distinctively chartered: For instance, being in a college preparatory curriculum (in contrast to a vocational one) makes a considerable difference in the aspirations and expectations of American high school students (Alexander and Eckland, 1975; see also Rosenbaum, 1975).

Intervening variables in all these processes, of course, would include not only the expectations of students themselves, but also those of their parents, teachers, counselors, peers, and peer cultures.

Proposition 6--the idea that education affects the qualities of people by allocating them into roles which are elaborated and high in status--explains a second major paradoxical finding in the current sociology of education. The direct long-run effects of schools on persons are thought to be rather moderate: But surveys of adults on almost any dependent variables--attitudes, values, information, or participation--almost uniformly show that education plays a dominant role. The argument here is that correlations between education and personal qualities are increased--or at least maintained--by a structure of subsequent allocation which provides distinctive life experiences for the educated. For instance, education can affect a person's sense of political efficacy by making him a powerful civil servant as well as by socializing him to a civic culture.

Research Implications of Allocation Theory: If taken seriously, and not simply used as a cynical critique of education, allocation theory would completely reorganize current research styles in the sociology of education. Allocation rules, unlike simple socialization effects, reign over both the students and the non-students, the educated and the uneducated, the graduates and those who never attended.

Research implication 1: effects on non-students:

Hypothesis 1: The creation of social rules allocating status and competence to graduates leads to increases in the socialization of students.

Hypothesis 2: But such institutions should lower the prospects of non-students, and in a sense desocialize them.

The more binding the allocation rules, the earlier and more convincingly are non-students committed to passive roles in society. This means that the credentialled society could well lower (below the previous floor) the modern competence of people of low education. Comparative contextual research is required to test this idea, since the independent variable is a property of the social system.

This argument has it that in a modern society education allocates its dropouts to failure. They (and their parents and friends) anticipate and adapt to this.

> Hypothesis 3: Similarly, subsequent to the period of schooling, non-graduates are socialized through life-experiences to the meaning of their failure just as graduates are socialized to the meaning of their success. The lagged differentials created by education should be greater, the more firmly the principle of educational allocation is established. Again, comparative research is required.

> Hypothesis 4: Those admitted to chartered educational organizations find their prospects enhanced even before attendance, while those rejected find their prospects lowered. They adapt their personal qualities in anticipation, even prior to attendance. These differentials should be greater, the stronger the allocative position of the school. For example, Benitez (1973) finds that students admitted to a national elite high school in the Phillippines gain in self esteem and "competence" even before their socialization begins. Wallace's (1966) data suggest a similar interpretation.

Effects such as these should help distinguish allocation theory from traditional socialization ideas.

Research implication 2: aggregate effects: A major implication of allocation theory is that inferences to the aggregate effects of education made from individual data on the basis of traditional socialization ideas are almost completely illegitimate. Researchers in the economics of education conventionally infer aggregate economic effects of education from income differentials between the educated and the less educated (see, e.g., the papers in Blaug, ed. 1968, 1969). It is assumed that these income differentials reflect real added value--the socialization gains of the educated. But if education is simply an allocation system, the gains of the educated may simply occur with equivalent losses for the uneducated. The expansion of education and educational allocation may have no effect on the aggregate product at all (Collins, 1971).

Similarly researchers on the political effects of education often infer that, because the educated occupy politically central positions, education must have helped create these positions (see, e.g., the papers in Coleman, ed., 1965; or the review by Meyer and Rubinson, 1975). But if education is simply a system of allocation, huge positional and attitudinal differences between the educated and the uneducated may exist with no aggregate effect at all. Igra, (1976), in fact, shows (using Inkeles' data) that increases in the aggregate development of societies lower the political participation of individuals of given education (though the political information of individuals is found to be enhanced). Such "frog-pond effects" at lower levels of analysis are discussed by Davis (1965), Meyer (1970), and Alexander and Eckland (1975).

The main arguments of allocation theory are added to those of socialization theory in Figure 2. Allocation ideas are discussed with some frequency in the current literature, though their research implications remain only partially explored.

The main problem with the theory, however, lies in its cynicism. Why do modern societies so consistently create and institutionalize huge systems of education, if these systems are simply expensive social sorting devices (Spence, 1973; Thurow, 1975)? That is, where do the basic allocation rules which are central to the theory (and modern societies) come from? At this point, even the cynics speak of education as in some way legitimating the modern class structure (Jencks, 1972; Bowles and Grintis, 1976). But this idea, its meaning, and its research implications, remain simply an indictment of modern society (e.g., Illich, 1971). Modern societies are "schooled". What does this mean?

LEGITIMATION THEORY

Modern extended and institutionalized systems of education build into society certain rules which actors take for granted, know others take for granted, and incorporate in their decisions and actions.* Thus institutionalized educational systems create a situation in which social gatekeepers (e.g., personnel officers)--even if they read and believe Ivar Berg's book--nevertheless know that they must hire people on the basis of educational credentials.

Two closely related aspects of modern educational systems are relevant here as independent variables: (1) they are extended as systems of classification, categorizing entire adult populations by level and specialty; and (2) they are institutionalized, with their classifications often controlled by the state and enforced in daily life by credentialling rules written into law and organizational practice. Almost everywhere, education is made compulsory and universal by national law, [often in the national constitution (Boli-Bennett, 1976)], and in most countries its structure is closely regulated by the nation-state (Ramirez, 1973; Rubinson, 1973).

---

*Actors may also internalize these rules as personal commitments, but this is less important--the critical aspect is that they internalize them as social facts and social realities (institutions which rely on personal beliefs, or even permit the question of personal beliefs to be relevant in social action, are less highly legitimated in important senses, than those which operate as realities).

Why does this occur? Whatever the economic origins of the process, the fact that it is usually accomplished and regulated by the state--unlike many aspects of economic development, which are left to individuals and subunits--suggests that its immediate origins lie in the political system: society as corporate organization (Swanson, 1971) rather than as a system of exchange. Formalized educational systems are, in fact, theories of socialization institutionalized as rules at the collective level. The three core propositions we used earlier to summarize traditional socialization theory become the structural basis of the educational system: Proposition 1--the idea that the schools teach critical skills and values--becomes institutionalized as the basic educational classification system: Education proceeds in a sequence (irreversible by ascriptive definition) from kindergarten through post-doctoral study, and covers a defined series of valued substantive topics. The student is a "high school graduate" and has had compulsory units of history and English and mathematics. It is an institutionalized doctrine, since for many purposes one must treat the student as having acquired this knowledge by virtue of the units or credits completed, not by direct inspection. Proposition 2--the idea that schooled qualities are carried into adult effectiveness--is institutionalized in the basic credentialling rules which dominate personnel allocation in modern society. If one hires an executive, a civil servant or a teacher, one must inspect educational credentials--it is optional whether one inspects the person's competence. A teacher or a doctor who graduated from school in 1930 is still a socially and legally valid teacher or doctor. Proposition 3--the idea that educational allocation creates social progress--is institutionally embedded in our doctrines of progress: It consists of modernity, professionalization, and rationalization. The possession of the best certified and educated people is a main index of the advanced status of a hospital, a school, often a business organization, and indeed a society itself.

Educational systems themselves are thus, in a sense, ideologies. They rationalize in modern terms, and remove from sacred and primordial explanations, the nature and organization of personnel and knowledge in modern society. They are, presumably, the effects of the reorganization of modern society around secular individualism which is a main theme of Marx and Weber. Our problem here, however, is to discuss their effects.*

Legitimating Effects of Expanded and Institutionalized Education

We discuss legitimating effects of education in four general categories created by the intersection of two dichotomies. First, education functions in society as a legitimating theory of knowledge defining certain types of knowledge as extant and as authoritative. It also functions as a theory of personnel, defining categories of persons who are to be treated as possessing these bodies of knowledge and forms of authority.

---

*The discussion which follows deals exclusively with the effects of institutionalized education on other aspects of society. Obviously, important causal effects also run the other way (see Meyer & Rubinson, 1975, for a review). Empirically disentangling the reciprocal effects requires data on societies over time.

Secondly, education validates both elites and citizens. Discussions of the legitimating function of education often emphasize only its role in supporting elites and inequality (e.g., Bowles and Gintis, 1976; Carnoy, 1972). But the overwhelming kind of education in the modern world is mass education(Coombs, 1968), closely tied to the modern state and notion of universal citizenship (Marshall, 1948; Bendix, 1964; Habermas, 1962).

These two distinctions define four types of legitimating effects of education, as specified in Figure 3. We discuss them in turn.

1. The authority of specialized competence. Education does not simply allocate people to a fixed, zero-sum, set of positions in society. It expands the authoritative culture and the set of specialized social positions entailed by this culture. Thus the creation of academic economics means that new types of knowledge must be taken into account by responsible actors. The creation of psychiatry means that former mysteries must now be dealt with in the social organization. The creation of academic programs in business management brings arenas of decision-making from personal judgment, or luck, to the jurisdiction of rationalized knowledge. Social problems call for human relations professionals (occasionally even sociologists). Safety or environment problems call for industrial or environmental engineering.

The point here is that quite apart from the immediate efficacy of the bodies of knowledge, they are authoritative and must be taken into account by actors at risk of being judged negligent or irrational. The business manager who plans by the seat of his pants--unblessed by economic projections--has no excuse for ignoring the best advice. The political leader who sees social problems as beyond analysis or cure is reactionary and primevil. The emotionally disturbed person who rejects psychiatry is displaying irrationality.

Thus the knowledge categories of the educational system authoritatively enter into daily life. Mysteries are rationalized, brought under symbolic control, and incorporated into the social system. Society and its subunits are buffered from uncertainty (Thompson, 1967). Thus:

> Proposition 7: The expansion (and institutionalization) of education expands the number of functions brought under social control.*

2. Elite certification: The source of modern allocation rules, which are central to allocation theory but unexplained by it, now becomes clearer. These rules are rooted in the institutionalization of education as a personnel theory in modern society. They justify or legitimate in modern cultural terms the specialization and stratification of modern elites:

> Proposition 8: The expansion (and institutionalization) of education creates and intensifies rules of allocation linking education to adult statuses.

Education, that is, not only creates economic knowledge which must be taken into account by rational actors. It is an initiation ceremony designating who authoritatively possesses this knowledge. So it also creates, not only psychiatry, but psychiatrists; not only modern management ideology, but M.B.A.s. The rational actor must take into account medical knowledge; and to do so he must consult a doctor. Thus, the modern organizational structure of society incorporates legitimated bodies of knowledge by incorporating the designated personnel.

---

*This assertion, incidentally, formalizes an idea of Schumpeter (1950, Ch. 12) about the way in which the intellectual optimism of modern capitalistic society generates its own institutionalization and destruction. The intellectuals rationalize more and more social functions, which are then brought under collective social and political control and removed from the market.

We take too narrow a view if we see this process as involving only a few specialized occupations. The most important credentialling rules are more general: the set of rules which connect the educational status college graduate (and high school graduate) with all sorts of formal and informal elite positions. These rules define a generalized body of elite knowledge and specify its legitimate carriers.

It now becomes clear why views of educational allocation as "zero-sum"--allocating a fixed set of social statuses--are wrong. Education creates new bodies of knowledge and personnel which must be incorporated.

> Proposition 9: The expansion (and institutionalization) of education expands the number of specialized and elite positions in society.*

The point here is that institutionalized education does more than simply allocate some to success and others to failure. The educated learn to claim specialized functions and to legitimate the specialized functions of others. The less educated learn that they are part of a social world elaborated far beyond the traditional community. This is one of the core meanings of the modern social status citizen.

---

*This proposition is impossible in conceptions of social status as simply a rank position, and thus as zero-sum. But there is no reason to assume that the total amount of status (or for that matter power) in society is fixed. Independent of their ranks, statuses (and whole status distributions) may vary in the expansion of their substantive rights and powers. It has often been observed, for instance, that American society--with a rank distribution very similar to many others--attaches fewer diffuse status prerogatives to each status in the system, and hence that money assumes special importance. We argue, above, that education expands the status rights attached to many positions in society, without necessarily altering the rank structure. This conception of status reflects Weber's original formulation.

3. The universality of collective reality: Mass education creates a whole series of social assumptions about the common culture of society, and thus expands the social meaning of citizenship, personhood, and individuality (modern ideas, all). It establishes a whole series of common elements for everyone:*

(a) It creates the assumption of a national language or languages and defines universal literacy;

(b) It reifies a given national history;

(c) It constructs a common civic order--common heroes and villians, a common constitutional and political order with some shared cultural symbols and with legitimate national participation;

(d) It validates the existence of a common natural reality through science and a common logical structure through mathematics, and in this way constructs a myth of a common culture intimately linked to world society; and

(e) It constructs broad definitions of citizen and human rights as part of the modern world view.

---

*We provide, here, a conventional list of the putative effects of mass education. But our argument is that whether or not these effects really occur on students, they enter into social life as taken for granted assumptions. Many Americans are not literate in the national language. But we treat each other, expect elites to treat us, and organize our public life, as if we all are.

Regardless of what people actually learn in school about their language and culture, nationally institutionalized mass education creates the assumptions of a national culture. For many purposes, both elite and citizen actors must take them into account.

> Proposition 10: The expansion (and institutionalization) of education expands the content and jurisdiction of the elements taken for granted as part of collective reality.

4. The extension of citizenship: Beyond defining and extending national culture, mass education defines almost the entire population as possessing this culture, as imbued with its meanings, and as having the rights implied by it. Mass education defines and builds the nation (Marshall, 1948; Bendix, 1964). It establishes persons as members of the nation over and above subgroups. It creates a public, and hence, ideas of public opinion as a vital force in society (Habermas, 1962; Bergesen, 1977). Individuals are now defined as possessing the competencies and the moral orientations to participate meaningfully in collective life.

> Proposition 11: Mass education expands the number of persons seen as possessing human and citizenship capacities and rights.

Research Hypotheses in Legitimation Theory

Ideas about institutions and about legitimacy often seem vague or platitudinous in our individualist and reductionist language. We therefore need to spell out the meaning of the propositions above in research terms.

Legitimation theory argues for the effects of the extension and institutionalization of national educational systems on the judgments, perceived realities, and actions of given individuals--ordinary persons, rule makers and critical social gatekeepers. The education (and socialization) of these individuals is to be held constant. Ideally, the actual distribution of socialized qualities in the societies being studied would be held constant too; for we are discussing the effects of education as an instituionalized system, not as the aggregate of some individual experiences. Thus, the research question is: How do the expanded distributions of schooled individuals, and the national institutional controls over the meanings of these distributions, affect the judgments and actions of individuals when their own education and the actual socialization of the population is held constant.*

Thus, if Propositions 7--9 are true, people of given education should see more education-based elites as having more authority in more substantive areas in societies with expanded and highly institutionalized educational systems. This should be true even when the size or actual resources of a given elite group is held constant. Ordinary people, political and business leaders, even military leaders, should be more likely to believe that societal problems demand educated expertise, in such societies. Even if these people do not themselves believe in particular experts, they should imagine that others do so in more "schooled" societies, and thus that educated elites have authority.

Evidence relevant to these ideas could be obtained from contextual analyses of comparative survey research data. It would be difficult, however, to eliminate from such analyses the obvious alternative hypothesis that people respond differently to elites in societies with expanded and institutionalized education because such societies really have more competent elites.

---

*Of course, this is an extremely stringent test of legitimation ideas, since it seems clear that according to the theory itself, the actual socialization of the population will itself expand in response to educational institutionalization (that is, people will acquire the expanded role properties being legitimated).

Experimental studies could help deal with this problem: Subjects could be confronted with hypothetical societies in which to engage in business or political planning. Would they be more likely to propose to use economists, social scientists, or human relation personnel, in societies described as having highly institutionalized and prestigious academic fields in these areas (holding constant the descriptions of the number and actual skill of the specialists involved)? Perhaps it would be enough to simply describe a society in which elite education is highly developed and institutionalized.

In the same way, either survey or experimental studies could show that in societies with more expanded and institutionalized elite education, individuals tend to see more specialized social functions and problems as coming under the explicit jurisdiction of collective society, and especially the state.

More specific studies could trace the effects of particular forms of elite authority from their institutionalization in the educational system to their acquisition of a position of (a) strength and (b) monopoly in society.* For example, the rise and effects of programs in business management, criminology, or personnel psychology could be traced either with cross-national survey data or with experiments asking individuals to make choices in different hypothetical societies.

---

*Causal effects run the other way, too. Powers in society tend to root themselves in the educational system.

The same ideas apply to studies of mass education. Propositions 10 and 11 argue that persons (ordinary people, economic elites, political elites, and colonels contemplating coups) would be more likely to see citizens as possessing expanded rights, competencies, and levels of national culture, in a society with expanded mass educational institutions. Thus ordinary persons and politicians would be likely to perceive national loyalty and widespread literacy in a society with universal mass education (even with actual loyalty and literacy held constant). Employers would be more likely, in such a society, to see action around an impersonal labor market as appropriate. Such questions could be investigated either with comparative survey data or experimentally: Samples of colonels (or revolutionaries) could be asked to plan coups in hypothetical societies. Presumably, the more mass education is described as institutionalized, the more likely the subjects would be to propose coups organized around the manipulation of public opinion, and the more likely they would be to see the subsequent regime as organized around modern mobilizing strategies (whether left or right wing) as opposed to traditional authoritarianism. These effects should remain even if the levels of societal literacy described are held constant.

These research suggestions make clear the nature of legitimation effects: Modern educational systems formally reconstruct, reorganize, and expand the socially-defined categories of personnel and of knowledge in society. They expand and rationalize these social realities as they enter the choices of the socialized and the unsocialized, the allocated and the unallocated. Education is, as has often been noted, a secular religion in modern societies: It provides a legitimating account of the competence of citizens, the authority of elites, and the sources of the adequacy of the social system to rationally maintain itself in the face of uncertainty.*

The Impact of Educational Legitimation on Allocation and Socialization

Education legitimates the personnel and context of modern role structures, making the allocation of people and authority appear rational and just. This greatly intensifies and transforms the operation of allocation and socialization theories:

> Proposition 12: The more institutionalized the modern system of education, the more intensified are the causal relationships of allocation and socialization.

We consider the particular instances of the generalization in turn; reviewing our earlier propositions summarizing allocation and socialization ideas:

1. As noted above (in Propositions 8 and 9) institutionalized education creates and intensifies educational allocation rules, and transforms allocation from a zero-sum process to a process producing new roles and meanings. Some people become "doctors": but the others don't become "non-doctors": they become patients. Education expands roles and sets them into proper relations with the rest of society.

---

*Modern education not only expands each society structurally: It brings societies into closer organizational similarity with each other. Societies come to be made up of more and more similar elites--often in professional communication with each other--and masses with more and more shared social rights. This organizational homogeniety means that information--and exploitation--can proceed very rapidly. New ideas and techniques are not alien. They are the stock in trade of an already incorporated profession and can thus be adopted with less resistance. So, in the modern world, the presence of locally-controlled, but organizationally similar educational systems in almost all nation-states makes possible the rapid cultural penetration of techniques (and political revolution). And it makes possible new kinds of dependence (e.g., a "brain drain").

2. The consequences of allocation, too, are intensified. The lagged socialization of the allocated (Proposition 6) is not simply an adaptation to their power, but is an affirmation of their authority. The process by which students in school acquire chartered qualities (Proposition 5) takes on additional meaning because of its legitimacy. The students (and non-students) are learning more than their own futures. They are also learning that the practical categories and topics of education give legitimate meaning to these futures. For instance, the college student learns something of a subject--say, sociology--because he is taught it (socialization) and because he and others expect and will expect graduates to know a bit about it (chartering). These processes are intensified by the legitimating reality of sociology: Both graduate and non-graduate students learn that it exists as a body of knowledge, and an allocative principle. This awareness intensifies the chartering impact on students: They acquire their sociology with a dutiful passivity which reflects their understanding that the status of this body of knowledge exists beyond their own acceptance or rejection. Whether or not they accept this discipline, their socially valid degree will reflect so many units of sociology.

3. Socialization processes are similarly intensified by the legitimacy of educational authority. Students experience an organization which has power over them, but which also controls core meanings and values in society. This moral authority undoubtedly intensifies socialization over and above that found in routine training organizations (Bidwell & Vreeland, 1963). Graduates carry through their lives these same meanings, not just their practical advantages. The autocorrelations in educational skills and values are probably increased by the fact that

these qualities are central parts of an individual's moral biography.

The Impact of Educational Legitimation on Educational Organizations

The legitimating effect of educational organizations--but also much of their socializing and allocating power--is derived from their highly institutionalized status in society. Operating at the institutional level as an authoritative theory of personnel and knowledge in society, the schools constitute a crucial ritual system: a system of initiation ceremonies (personnel) and of classifications of information (knowledge).

This makes it clear why schools often seem to act as ritual organizations, sacrificing "effectiveness" for classificatory rigidity (Meyer & Rowan, 1975; Kamens, 1977). Their larger social effectiveness (and their claim to resources) inheres precisely in this ritual structure: the apparatus of classes and levels and degrees and subjects. By emphasizing their formal ritual structures, schools maximize their links to their main source of authority, their main resources, and quite possibly their main effectiveness. Dramatizing their structures as socially legitimated and legitimating initiation ceremonies informs students (and others) both about the payoffs to which they can adapt, and about the fact that those payoffs are highly proper, deriving from the core meaning and values of society. Ritualism, thus, by Proposition 12, reinforces the immediate effectiveness of schools in dealing with students.

Summary of legitimation theory: Legitimation theory suggests two general ideas on school effects: Institutionalized education, as a theory of personnel and knowledge, affects society directly, apart from the training and allocation of students. Institutionalized education creates and intensifies the relationships of socialization and allocation. These two main themes are added to the explanatory structure presented earlier in Figure 3.

CONCLUSION

Schools may teach people useful skills and values. Whether they do or not in particular cases, they certainly allocate people to higher social status positions; and this affects the anticipations and socialization of the students (and non-students) as well as the experience and later socialization of the graduates (and non-graduates).

The allocating power of the schools, in turn, derives from their status as social institutions creating and validating categories of personnel and knowledge. The schools enhance the number and legitimacy of these categories--far beyond levels possible with more primordial myths of the origins of personnel and knowledge--and thus expand the whole rationalized modern social structure. These legitimating effects of schools reconstruct reality for everyone--the schooled and the non-schooled alike. They also intensify the meaning and validity of allocating and socializing processes.

So a student is in a position of experiencing (a) the immediate socializing organization, (b) the fact that this organization has the allocating power to confer status on him, and (c) the broader fact that this allocation power has the highest level of legitimacy in society. The education he receives has a very special status and authority: Its levels and content categories have the power to legitimately redefine him in the eyes of everyone around him, and thus take on overwhelming ceremonial significance.

Research on such questions must examine the effects of education as an institution, considering effects of variables quite beyond the level of the classroom, the peer group, or the school as an organization. Either experimentally, or with cross-societal (or time series) analyses, we need to consider the contextual effects of variations in the extension and institutionalization of education on the perspectives of students and non-students, graduates and non-graduates, citizens and elites. If education is a myth in modern society, it a powerful one. The effects of myths inhere, not in the fact that individuals believe them, but in the fact that they "know" everyone else does, and thus that "for all practical purposes" the myths are true. We may all gossip privately about the uselessness of education, but in hiring and promoting, in consulting the various magi of our time, and in ordering our lives around contemporary rationality, we carry out our parts in a drama in which education is authority.

Figure 1

Traditional Socialization Theory

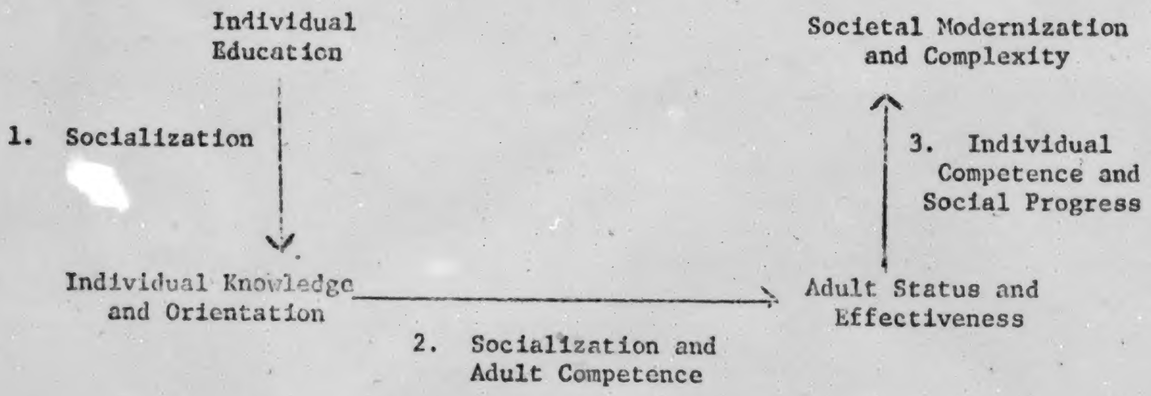
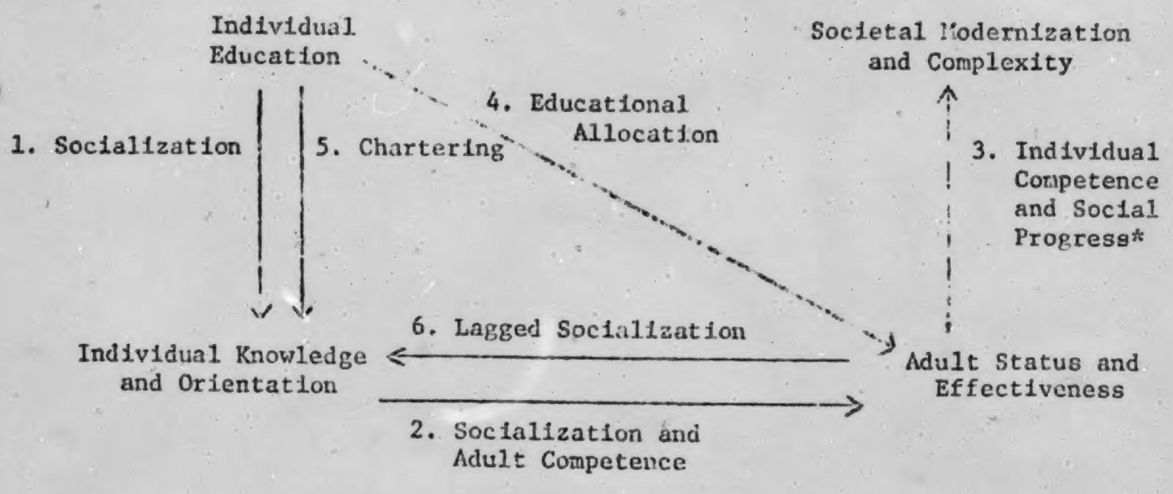

Individual Education
1. Socialization
Societal Modernization and Complexity
3. Individual Competence and Social Progress
Individual Knowledge and Orientation
Adult Status and Effectiveness
2. Socialization and Adult Competence

Figure 2

Allocation Theory Combined with Socialization Theory


Individual Education
Societal Modernization and Complexity
4. Educational Allocation
1. Socialization
5. Chartering
3. Individual Competence and Social Progress*
6. Lagged Socialization
Individual Knowledge and Orientation
Adult Status and Effectiveness
2. Socialization and Adult Competence

*Pure allocation theory suggests that this effect is irrelevant. If a given set of adult competencies are simply allocated by education, no net societal gain in number of competent individuals will occur.

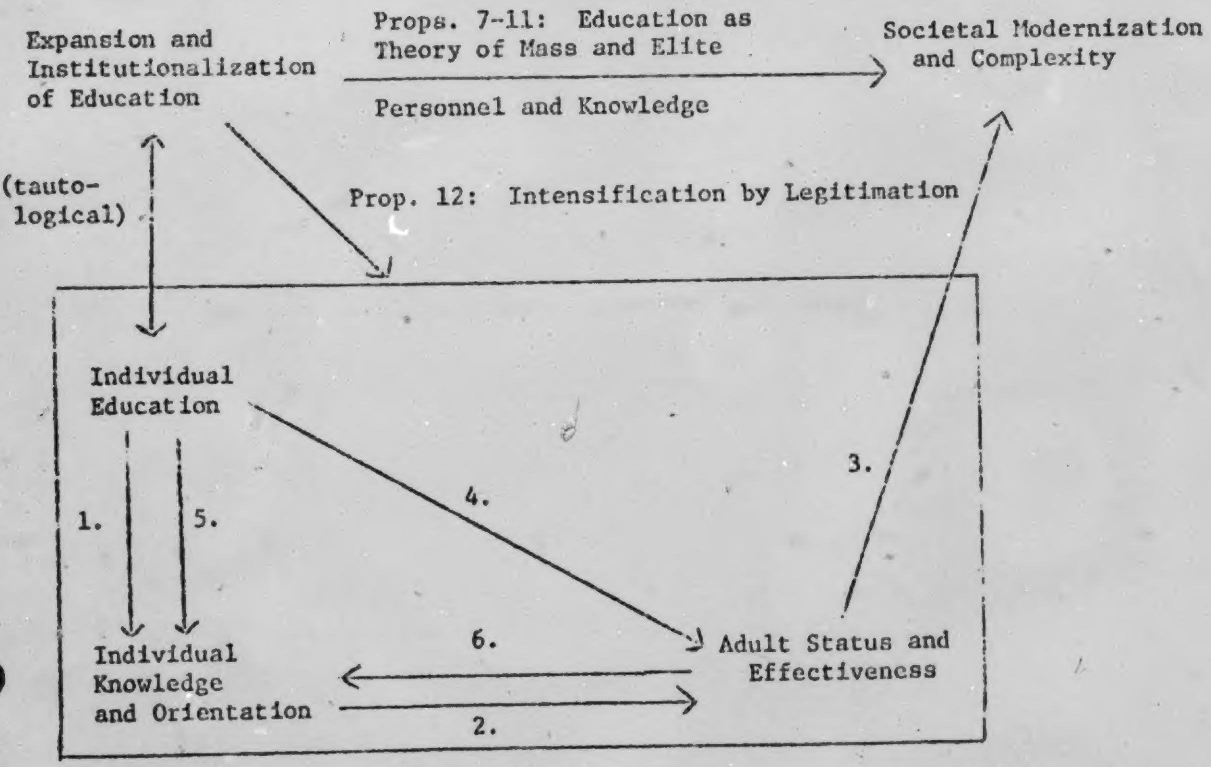
Figure 3

Types of Legitimating Effects of Education

	Elite Education	Mass Education
Education as Theory of Personnel	1. The Authority of Specialized Competence	3. The Universality of Collective Reality
Education as Theory of Knowledge	2. Elite Certification	4. The Extension of Membership: Nation-building and Citizenship

Figure 4

Legitimation Effects of Education


Expansion and Institutionalization of Education
Props. 7-11: Education as Theory of Mass and Elite
Personnel and Knowledge
Societal Modernization and Complexity
(tauto-logical)
Prop. 12: Intensification by Legitimation
Individual Education
1.
5.
4.
3.
Individual Knowledge and Orientation
6.
2.
Adult Status and Effectiveness

References

Alexander, K. & Eckland, B. Contextual effects in the high school attainment process. American Sociological Review, 1975, 40, 402-416.

Almond, G., & Verba, S. The civic culture. Princeton: Princeton, University Press, 1963.

Astin, A. & Panos, R. The educational and vocational development of college students. Washington: American Council on Education, 1969.

Bendix, R. Nation-building and citizenship. New York: John Wiley, 1964.

Benitez, J. The effect of elite recruitment and training on diffuse socialization outcomes. Unpublished doctoral dissertation, Stanford University, 1973.

Berg, I. Education & jobs. Boston: Beacon Press, 1971.

Bergesen, A. Political witchhunts: A cross-national analysis. American Sociological Review, 42, 1977.

Bidwell, C. & Vreeland, R. College education and moral orientations. Administrative Science Quarterly, 8, 1963, 166-191.

Blau, P.M. & Duncan, O.D. The American occupational structure. New York: John Wiley, 1967.

Blaug, M. (Ed.), Economics of education. Baltimore: Penguin Books, 1968, 1969.

Boli-Bennett, J. The expansion of constitutional authority in nation-states, 1870-1970: A quantitative assessment. Presented at the annual meetings of the Pacific Sociological Association, San Diego, California, April 1976.

Bowles, S. & Gintis, H. Schooling in capitalist America. New York: Basic Books, 1976.

Carnoy, M. (Ed.), Schooling in a corporate society. New York: David McKay, 1972.

Coleman, J. (Ed.), Education and political development. Princeton: Princeton University Press, 1965.

Collins, R. Functional and conflict theories of educational stratification. American Sociological Review, 1971, 36 1002-1018.

Coombs, P. The world educational crisis. New York: Oxford University Press, 1968.

Davis, J.A. The campus as a frog pond. American Journal of Sociology, 1966, 72, 17-31.

Duncan, O.D., Featherman, D.L. & Duncan, B. Socioeconomic background and achievement. New York: Seminar Press, 1972.

Feldman, K.A., & Newcomb, T.H. The impact of college on students. San Francisco: Jossey-Bass, 1969.

Habermas, J. Structurwandel der Öffentlichkirt. Neuwied, 1962.

Himmelweit, H. & Swift, B. A model for the understanding of school as a socializing agent. In P. Mussen, et al. (Eds.), Trends and Issues in Developmental Psychology, New York, Holt, 1969.

Holsinger, D. Education and occupational attainment in Brazil. Comparative Education Review, 19, 1975.

Igra, A. Social mobilization, national context, and political participation. Unpublished doctoral dissertation, Stanford University, 1976.

Illich, I. Deschooling society. New York: Harper and Row, 1970.

Inkeles, A. & Smith, D.H. Becoming modern: Individual change in six developing countries. Cambridge, Harvard University Press, 1974.

Jencks, C. Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books, 1972.

Kamens, D. The college "charter" and college size: Effects on occupational choice and college attrition. Sociology of Education, 1971, 44, 270-296.

Kamens, D. Colleges and elite formation: The case of prestigious American colleges. Sociology of Education, 1974, 47, 354-378.

Kamens, D Legitimating myths and educational organizations. American Sociological Review, 42, 1977.

Kerckhoff, A.C. Patterns of educational attainment in Great Britain. American Journal of Sociology, 1975, 80, 1428-1437.

Marshall, T.H. Citizenship and social class. New York: Doubleday & Co., 1948.

Meyer, J. The charter: Conditions of diffuse socialization in schools. In W.R. Scott (Ed.), Social processes and social structures. New York: Holt, Rinehart and Winston, 1970. 564-578.

Meyer, J. High school effects on college intentions. American Journal of Sociology, 1970a, 76, 59-70.

Meyer, J. Theories of the effects of education on civic participation in developing societies. SEADAG Papers Series, New York, The Asia Society, 1973.

Meyer, J. & Rowan, B. Notes on the Structure of Educational Organizations. Annual meetings of the American Sociological Association, San Francisco, California, August 1975.

Meyer, J. & Rubinson, R. Education and political development. Review of Research in Education, 1975, 3, 134-162.

Mischel, W. Introduciton to Personality. New York: Holt, 1971.

Plant, W. Longitudinal changes in intolerance and authoritarianism for subjects differing in amount of college education over four years. Genetic Psychology Monographs, 72, 247-287.

Ramirez, F. Societal corporateness and the nationalization and expansion of educational systems. Paper presented at the meeting of the American Sociological Association, New York, August 1973.

Rosenbaum, J. The stratification of socialization processes. American Sociological Review, 1975, 40, 48-54.

Rubinson, R. The political construction of education: Conditions for the creation of membership in political society. Paper presented at the meeting of the American Sociological Association, New York, August 1973.

Schumpeter, J. Capitalism, socialism and democracy. New York: Harper & Row, 1950.

Spence, M. Job market signalling. Quarterly Journal of Economics, 87, 1973, 355-375.

Stinchcombe, A. Constructing social theories. New York: Harcourt, 1968.

Swanson, G. An organizational analysis of collectivities. American Sociological Review, 1971, 36, 607-623.

Thompson, J.D. Organizations in action. New York: McGraw-Hill, 1967.

Thurow, L. Generating inequality. New York: Basic Books, 1975.

Wallace, W.L. Student culture: Social structure and continuity in a liberal arts college. Chicago: Aldine, 1966.

Methodological Problems in Analyzing the Effects of Educational Organizations on Social Structure*

Michael T. Hannan
Stanford University

Technical Report No. 14

August, 1976

Consortium on Methodology for Aggregating
Data in Educational Research

*This research was conducted under National Institute of Education contract #NIE-C-74-0123

Vasquez Associates, Ltd.
1744 N. Farwell Ave.
Milwaukee, Wisconsin 53202

METHODOLOGICAL PROBLEMS IN ANALYZING THE EFFECTS OF EDUCATIONAL ORGANIZATIONS ON SOCIAL STRUCTURE

Michael T. Hannan
Stanford University

August 1976

Paper presented at conference on Polish and American Youth: Higher Education and Changes in Social Structure, Warsaw, sponsored by the Polish Academy of Sciences. This research was supported by NIE Research Contract NIE-C-74-0123.

My purpose in this paper is to identify a set of broad related obstacles to modeling the nature of the impact of education on social structure. In the most general sense each of the issues concerns difficulties in bounding social systems to study the consequences of educational institutions. Three specific instances of the problem are developed. The first concerns the problem of choosing the level of analysis at which to study impacts of educational organizations and institutions. For what purposes is it appropriate to study only the impacts on individuals obtaining education and for what other purposes must we consider broader outcomes, (e.g. the effects of the expansion of schooling on those who do not attend school)? In discussing the issues involved I concentrate on two substantive examples, studies of the impact of higher education on productivity and politics. The second specific issue concerns bounding processes in time. How do we go about choosing a time frame within which to study the impacts of education; and what difference does choice of time perspective make? The third and final issue concerns problems that arise in analyzing the effects of educational organizational structure on outcomes of students, (e.g. academic performance, political socialization, etc). Most educational systems are hierarchically organized (e.g. study programs within academic departments, departments within schools, schools within universities, universities within national educational organization). Organizational reforms are introduced at every level and may have had consequences at every other level. Methods of analysis currently employed by sociologists do not do a good job of dealing with hierarchical structure in evaluating consequences of structural variations. I propose an alternative approach.

Sociological methodology has been greatly affected over the past ten years by large scale infusions of econometrics (cf. Blalock 1971; Goldberger and Duncan 1973). The main emphasis in the so-called structural equation tradition is on the need for proper specification of causal relations among variables prior to empirical analysis. The resulting emphasis on modeling has proven highly beneficial to some areas of sociology. Without diminishing at all the value of this recent work let me suggest that the single-minded emphasis on a priori statement of causal structure invites deemphasis of several other critical features of the research process. Here I focus on one of these relatively neglected issues: Bounding systems under study.

An accurate causal depiction of the system of relationships among a set of variables requires at least two things. The first is specification of which variables causally affect others (as in the path diagrams so common in recent empirical reports in the sociological journals). A second element, usually implicit in sociological research, is a description of the state space. The state space of any process is the set of all outcomes that are relevant to judgements about the process. That is, it is an exhaustive enumeration of the possibilities facing the analyst together with an accounting of how each outcome is to be related to inferences about the nature of the process. The state description of a process includes specification of system boundaries (this is implied by the stipulation that the enumeration of outcomes be exhaustive) and about proper levels of analysis at which to observe the process. Moreover, if temporal considerations arise in the substantive problem, an adequate state description of a process must include temporal orderings of states in the system.

Failure to properly define the state space of a process impedes the production of consistent and meaningful empirical results even when models are properly specified with respect to which variables affect which other variables. Unless we understand this we will have great difficulty accounting for the lack of agreement among so many empirical studies using similar causal models. The theme of this paper is that inadequate attention to proper state description of processes relating educational organizations and institutions to changes in social structure is a major obstacle to the development of theory and research on the sociology of educational institutions.

Treatments of sociological methodology do not typically formulate issues in these terms. Therefore it is best to leave this high level of abstraction and shift to formulating the issues in the language of contemporary research practice. I will attempt to do this by considering three issues, each of which involves an instance of the general problem of consequences for empirical analysis of inadequate state space specification.

It will facilitate exposition to concentrate on substantive examples. Much of the social science literature on the effects of the expansion of higher education on social structure concentrates on economic or production outcomes. In the broadest terms this research explores whether the expansion of higher education in some social system increase productivity (e.g. produce increased levels of economic growth). Much of the economic research on this question analyzes the effects of increased years of schooling on individual wage rates (or incomes).[1] Such research is motivated by a human capital perspective that holds both that schooling is an investment in an individual's store of "capital", and that productivity and wage rates are proportional to capital stocks. If individuals with more education earn more than otherwise similar individuals, researchers in this tradition conclude that education does indeed improve capital stocks for individuals; and they add that such effects when aggregated imply that increases in the number of individuals attaining any level of schooling will lead to system wide gains in productivity.

There are at least two major flaws in this reasoning. First, suppose that the hypothesized relationship between schooling and wage rates holds. Then it is true at the individual level that schooling increases wage rates. But, wage rates may not reflect productivity; they may reflect statuses that are either institutionalized or created by education. More important for present purposes, there is a fallacy of composition. Just because the relationship holds at one level is no guarantee that it will hold at some higher level of analysis. Individuals with higher than average levels of education may do well in wage competition simply because education is used as a signal of certain competences (e.g. literacy, ability to cope with certain types of formal bureacracies). But if the educational system were expanded so that all individuals achieved some high level of education, the signal value of education would vanish and so too would the relationship between schooling and wage rates.

The second flaw becomes apparant when we consider the possible failure of the empirical relationship. Suppose there is no relationship between educational attainment and wage rates, or even actual productivity. Does this mean that increases in education do not affect levels of productivity in the social system? It does not. To see this consider the information produced by higher educational systems. Besides preparing individual students for their careers, universities contribute to the growth of science and technology. There is widespread agreement among historians and social scientists that improvements in technology account for most of the dramatic increases in productivity yet witnessed. So to the extent that systems of higher education contribute to technological growth, they affect economic growth. This is most true when there are strong informational barriers at the system boundaries. Then the growth in productivity of any system depends on the information processing capacity of that system and university structures can have maximum impact within the local system.

Both flaws arise due to an inappropriate state description of the process. The research design is focused at the individual level in a situation in which there are possible system level effects (that is effects that hold net of the composition of the composition of lower order effects). The impacts of education on economic growth are not bounded in any meaningful sense by the population that receives formal instruction in schools.

One of these criticisms can be leveled at my own research on the impact of education (Hannan 1973; Meyer et al. 1976; Hannan and Nielsen

1976). Our group has conducted panel analysis of a large number of nations to determine the impact over periods ranging from five to twenty years of the expansion of various levels of education on economic growth. We find that both primary and secondary education have strong and persistent effects; but we find no such effects for higher education. Are we to conclude that higher education does not affect national economic growth? Again I think the answer is no. The complication is one that I alluded to above. The information barriers around nations are relatively weak especially with respect to flows of scientific and much technical information. Moreover, there is much movement of educated persons among nations (the so called "brain drain"). Both of these situations may lead to investments in higher education in one nation having payoffs in many nations. If so, we should not find it surprising that we cannot find any clear relationship within nation-states. In other words, the processes by which university social organization affects economic growth is not bounded in any meaningful sense by national boundaries.[2]

These issues are not peculiar to research on the economic consequences of education. Consider research on the political consequences of expanded higher educational systems. There is ample evidence that exposure to higher education affects individual political behavior and attitudes (cf. Feldman and Newcomb 1969). But the system level consequences of increasing the proportions of the polity with higher educational degrees is not clear. It is not inconceivable that the expansion of university enrollments may have effects on those who do not attend universities that are as large as the impacts on university students.

This may happen whenever a sufficient mass of university graduates exists to change the structure of opportunity for political action for non-graduates. For example, in the United States access to positions of high political power is increasingly limited to those with degrees in law (there is also an apparant shift as well towards those with degrees in economics or other social sciences). Such a situation can become institutionalized only when there is a sufficient mass of degree holders. Once institutionalized, this new basis of access to power may produce significant changes in the types of political control exercised in the system (e.g. control by technical expertise). In other words, the expansion of an elite certifying institution of a certain form may result in changes in the nature of political action in the system quite apart from affecting the behavior of its graduates.

Once universities obtain substantial control of the elite certification process, control of universities becomes an essential resource to groups contending for power. It is also quite likely that universities will become symbolic embodiments of political movements. Both of these processes are readily apparant in multi-ethnic nations. A particularly clear example is the manner in which the Belgian linguistic conflict "settled" on the dispute over the language situation within the university at Louven (or Louvain -- which is what the fight was all about). My point is that as long as universities educate only or mostly the children of the traditional ruling class, their action does not much affect politics (though the politics of their graduates certainly does). But once universities take on the expanded role of educating the middle classes and the working class, the universities' political role maybe fundamentally altered. The manner in which the process of expansion occurs may both affect politics and become a central political issue. To attempt to explain the political importance of modern expanded universities only in terms of the political effects on students will miss the point almost completely. Analyses of the process that employ a state space description framed in these terms will not produce consistent and meaningful results.

It is well and good to urge that social scientists employ proper state spaces along with proper causal models in analyzing the effects of higher education on social structure. Perhaps all that is required is a more thoughtful treatment of this phase of research design. Nonetheless it would be valuable to have more explicit guidelines than now exist in the literature on sociological methodology. This is not the place to attempt this. So I will make only a few additional comments.

Most research on the system-wide consequences of education take the following form. A relationship is postulated at the individual level:

$$y = f(x)$$

where y is some outcome of interest such as some political behavior and x is some measure of exposure to higher education. Then the same relationship with the same functional form is postualted at the system level:

$$Y = f(X)$$

where Y and X denote values of y and x aggregated across the system:

$$X = g(x_1,\ldots,x_N)$$

$$Y = h(y_1,\ldots,y_N)$$

Two simplifying assumptions must hold for the state space description in terms of individual characteristics to be useful: (1) the relationships between individual values of variables and system level variables (e.g. between individual productivity and system productivity) must take a very simple form; and (2) the functional relations f and F must be identical (cf. Hannan 1971). If both of these conditions hold, that is if the laws of composition are simple and there are no direct system level effects net of the lower order (individual level) effects, one loses little by utilizing only a microscopic description of the system and the structure of causal relations.

I have suggested that the two crucial simplifying assumptions are unlikely to hold in most research on the effects of higher education on social structure. As is well known, national productivity is no simple function of the productivities of all of its members. Neither are national political outcomes simple functions of the political attitudes and behaviors of the members of the polity. Just as important, the laws relating variables are ordinarily different at the two levels. For example, this is true whenever the distribution of x in the population affects Y. I have suggested the operation of such "mass effects" in considering the political struggles for control over universities and the effects of such struggles on ethnic nationalist movements. Whenever there are such mass or threshold effects, microscopic state descriptions do not suffice. I have also suggested system level effects that do not depend upon individual x's and y's, (e.g. the effects of the institutionalization of higher education on elite certification on the nature of political control). Whenever direct system level effects operate, it is also obvious that microscopic state descriptions are not appropriate.

Because this conference focuses on impacts on social structure, I have emphasized the fallacy of composition which involves utilizing a state space description and a causal structure that is at too low a level of analysis. Of course the opposite error is at least as common in empirical research, reasoning from relationships holding among aggregates to relationships at the individual level (Hannan 1971). Each of the issues discussed above holds with equal force in this case, though the details of the problem differ.

The second instance of the general problem I wish to address concerns temporal descriptions of the causal system. For studies of the effects of educational organizations and institutions on social structure is it enough to employ static state descriptions, or should we introduce dynamic considerations? As Coleman (1968) has shown so clearly the answer to this question turns on the nature of the equilibria under study and on the nature of the equilibria seeking processes. Static (cross-sectional) analysis depends implicitly on either of two assumptions. Either the system under study is in or very near equilibrium or that the process of shifting from one equilibrium to another (when the system is shocked) is instantaneous (relative to our ability to measure states). Therefore, the practical question concerns whether either of these conditions is likely to be met in the research on the effects of educational organizations on social structure.

The first possibility is that the social systems under study are in or near equilibrium (in the sense that all of the competing forces operating on the system are in balance so that the system will not change without further environmental change). While economists and human ecologist have had reasonably good success working with models that rely on this assumption, I do not think that many sociologists will find it attractive. When higher educational institutions expand, they do so at the expense of other institutions (since each system has finite resources, cf. Nielsen and Hannan 1976). The very fact that one institution can grow rapidly at the expense of other institutions shows that the system is undergoing structural change and is not likely to be in equilibrium. Moreover these changes take place in the context of rapid industrialization, urbanization, bureaucratization, and other fundamental shifts in patterns of social organization. Clearly these are not systems "at rest."

The second assumption concerns speed of adjustment. If the features of social structure of interest (e.g. the economy, the polity) adjust very quickly to change in educational systems, we lose very little by ignoring dynamic considerations. I have conducted research on the converse problem, the speed of adjustment of educational systems to changes in social structure (Nielsen and Hannan 1976; see also Freeman and Hannan 1975). We find that higher educational systems adjust quite slowly relative to primary and secondary systems. And, the speed of adjustment appears varied with the wealth or level of development of the social system. Higher educational systems in rich nations adjust enrollments more rapidly than do those in poorer or less developed nations. We investigated these dynamics to test certain hypotheses concerning the population ecology of organizations (Hannan and Freeman 1976). But the findings also indicate that static analyses of the same phenomena would likely be confounded by the ignored dynamics. Preliminary findings on the effects of expansion of education on economic growth (Hannan and Nielsen 1976) also show dynamic effects of the same sort. Each of these studies raises doubts that the social systems of interest adjust quickly enough that dynamics can be safely ignored.

So this is a second sense in which the state descriptions of systems commonly employed in sociological research on consequences of education are inappropriate. We urge that researchers consider the possible dynamic nature of the processes and adjust research designs accordingly (e.g. conduct panel analysis).

The third broad set of issues concerns analysis of the effects of educational organizations on the people they process. To this point I have emphasized the need to model system level outcomes directly. But achieving a complete modeling of the impact of educational organizations on society also requires research on consequences to individuals. As I noted earlier this is indeed the focus of work, American research on educational organizations. Each of the issues discussed above also hold for research with this focus. A slightly different problem concerns specifying state spaces appropriate for hierarchical structures and designing analysis procedures that fit such specifications.

As I indicated at the outset, most educational systems are hierarchically arranged (in the sense of a partial ordering). This aspect of structure is most visible in primary and secondary education organization. Students in the U.S. are grouped into classrooms within grade levels, within schools, within school districts, within state educational administrations. But universities also have hierarchical structures. They tend to be divided into schools, (e.g. law, medicine, arts and sciences, engineering, business, etc,). Schools are further subdivided into departments, (e.g. history, sociology, mathematics). Departments are often further subdivided into programs or curricula, (e.g. experimental, cognitive and abnormal psychology). The methodological challenge is to employ the multiple levels of analysis in modeling the impacts of organizational structure on student careers and outcomes.

Let me first pose the problem in terms more appropriate to primary and secondary educational research and later return to the analogues in research on university organization. Consider the problem of modeling the effects of school structure on student achievement. Suppose we are interested in the effects of some property of school structure, (e.g. centralization of curriculum student teacher ratio, etc.) on achievement. In the American context there is systematic sorting of families into school districts that produces a correlation of individual student attributes, (e.g. ability, socio-economic status, educational aspirations, etc.) with school characteristics. Therefore an adequate state description of the achievement process must contain both student characteristics and school characteristics.

The practical problem facing the researcher is as follows. He may analyze either individual level data, (e.g. regressing individual achievement on student characteristics and school characteristics[3]) or he may analyze school level averages, (e.g. regress average achievement on average individual characteristics and school properties). We have shown (Hannan et al. 1976) that in most realistic cases (i.e. when the models are not perfectly specified) results of the two different analysis will be quite dissimilar. This finding adds considerable importance to the choice between models.

Some investigators argue that when the question is posed at the school level, the school level regression (or, more correctly, the between-school regression) is most appropriate. A forceful statement of this position is made by Bidwell and Kasarda (1975, 1976). Others (cf. Hannan et al. 1976) point out that research on this issue has not specified the type of system level (school level in this case) processes discussed earlier. The causal arguments concern impacts on individual students which are composed (by a very simple function, that of addition) to school level outcomes. That is, in terms of the two-fold criteria suggested earlier the microscopic state description has not yet shown to be inappropriate.

To date the discussion of the issues involved does not appear to have been resolved. Each group holds to its initial position. Therefore, it is valuable to search for a modeling strategy that turns the issue of taste into an empirical problem. I wish to propose that the framework I have discussed elsewhere for multi-wave panel analysis, the "pooling of cross-sections and time series model" (Hannan and Young 1974; 1976) provides a means for doing just that.

It will simplify exposition to begin first with the framework as it applies to panel analysis and then draw the analogues to hierarchical structure analysis. Suppose we are analyzing the relationship of Y to X in a panel design in which we have multiple waves, say T of them, on N individuals. Then the simplest model we can employ is

$$Y_{it} = \alpha + \beta X_{it} + u_{it}. \qquad (i = 1,\ldots, N;\ t = 1,\ldots,T) \qquad (1)$$

where all observations are pooled into a single model.

Suppose a researcher analyzed such a data structure by means of ordinary least squares (OLS) regression. We would almost certainly object that the observations are not independent; the responses of the same individual at different periods are presumably all influenced by his/her unobserved but relatively constant characteristics (e.g. personality, history). OLS depends on the assumption that observations are independently distributed; efficiency is reduced by correlated observations and tests of significance and confidence intervals calculated by OLS formulas are biased (in the "liberal" direction) in such cases.

When dealing with data structures in which left out variables (summarized as disturbances) are correlated, we usually turn to generalized least squares (GLS) (Goldberger 1973). As the name suggests this method is a generalization of OLS that has OLS as a special case when the observations are indeed independent (in the sense that the disturbances are uncorrelated). To see how it handles the autocorrelation problem for panel analysis, we must state our model more explicitly. It will help to write the model in matrix form:

$$\begin{aligned}
&\underline{Y}_t = \underline{X}_t\,\underline{\gamma} + \underline{u}_t \\
&Y = (Y_{11},\ldots,Y_{it},\ldots,Y_{NT})' \\
&X = (X_{11},\ldots,X_{it},\ldots,X_{NT})' \\
&u = (u_{11},\ldots,u_{it},\ldots,u_{NT})' \\
&\gamma = (\alpha,\ \beta)'
\end{aligned} \tag{2}$$

and each element of $\underline{u}$ say u_{it} contains a systematic component μ_i and a random disturbance ν_{it}:

$$u_{it} = \mu_i + \nu_{it} \ .$$

Further,

$$\underline{Euu'} = \sigma^2 \begin{bmatrix} \underline{A} & 0 & . & . & . & 0 \\ 0 & \underline{A} & & & & 0 \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ 0 & 0 & . & . & . & \underline{A} \end{bmatrix} \text{, an (NT x NT) matrix ,}$$

where

$$\sigma^2 = \sigma^2_\mu + \sigma^2_\nu \ , \quad \rho = \sigma^2_\mu/(\sigma^2_\mu + \sigma^2_\nu) \text{ , and } \underline{A} = \begin{bmatrix} 1 & \rho & . & . & . & \rho \\ \rho & 1 & & & & \rho \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ \rho & \rho & . & . & . & 1 \end{bmatrix} \text{, a (T x T) matrix.}$$

The mechanism that produces autocorrelation is quite simple; it states that there are omitted variables that are constant for each individual and affect the dependent variable in the same way every period. Consequently all disturbances for each individual are correlated at the same level, call this ρ. Moreover to simplify the model we have assumed that every individual is characterized by the same correlational structure in the distrubances. This is the model that Nerlove (1971) (see also Hannan and Young 1976) applied to the panel analysis problem.

The GLS strategy is to find a matrix Ω that has the form that:

$$Y_t \ \Omega^{\frac{1}{2}} = X_t \ \Omega^{-\frac{1}{2}} \ \beta + u \ \Omega^{\frac{1}{2}} \qquad (3)$$

and

$$Eu \ \Omega \ u' = \sigma^2 \begin{bmatrix} 1 & 0 & . & . & . & 0 \\ 0 & 1 & . & . & . & 0 \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ 0 & . & . & . & . & 1 \end{bmatrix}$$

That is, the matrix has the form that once the model is transformed, the resulting disturbances have the form under which OLS is optional. Then, in a sense, one applies OLS to the transformed data. In fact the two steps are done simultaneously. The GLS estimator is defined as

$$\underline{\hat{\beta}}_{GLS} = (\underline{Q}_t'\underline{\Omega}^{-1}\underline{Q}_t)^{-1}\underline{Q}_t'\underline{\Omega}^{-1}\underline{y}_t, \qquad (4)$$

where

$$\underline{\Omega}^{-1} = \begin{bmatrix} \underline{A}^{-1} & 0 & \cdots & 0 \\ 0 & \underline{A}^{-1} & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \underline{A}^{-1} \end{bmatrix},$$

and (cf. Hannan and Young, 1974)

$$\underline{A}^{-1} = (1/\eta)(\underline{I}_T - \underline{\ell\ell}'/T) + (1/\xi)(\underline{\ell\ell}'/T),$$

where $\eta = (1 - \rho)$, $\xi = (1 - \rho) + T\rho$, and $\underline{\ell}$ is a (T x 1) vector of ones.

The form of the GLS transformation (4) can be intuitively motivated as follows. The peculiar feature of pooled models is the use of both cross-sectional (between-unit) and longitudinal (within-unit) variation to estimate causal parameters. The richness of the data presents an implicit choice: how to weight one type of variation relative to another. Generalized least squares uses ρ to weight the two types of information. To see this, consider the case where $\rho = 0$. Then $\underline{A}^{-\frac{1}{2}} = \underline{I}_T$ and observations are transformed in (3) by an identity transformation. GLS reduces to OLS where cross-sectional and time series variation are weighted proportionately to N and T (see Maddala, 1971). At the other extreme, when $\rho = 1$, $\underline{A}^{-1} = \underline{\ell\ell}'/T^2$. It is easy to show that this transformation averages observations over time for each unit. The result is a regression on grouped observations where all of the weight is placed on cross-sectional variation. In cases where ρ takes on a value $0 \leq \rho \leq 1$, GLS weights time series variation inversely to ρ. Such a weighting is appropriate since ρ measures "redundancy" in the time series. The more redundancy, the lower the weight attached to longtitudinal variation.

The gain in efficiency of GLS relative to OLS arises because GLS uses a weighting based on population or sample information rather than an arbitrary weighting (cf. Goldberger, 1973). Such a gain can be realized whenever ρ is known *a priori* or can be estimated from sample data in a consistent (or unbiased) manner. When GLS is applied using the simple form of estimation of $\hat{\rho}$ proposed by Nerlove, GLS has very good small sample properties (Nerlove 1971; Hannan and Young 1974, 1976).

Now that we have stated the technical details of the model we can return to the substantive problem. That concerns the choice among two models for the achievement study example. The first is an individual level analysis of all students that treats each student as an independent observation. The second alternative is a between-school regression that ignores within-school variations. The analogy between this situation and the multi-wave panel problem is immediate. Schools play the role of individuals in the panel design and students play the role of observations on individuals; that is individuals are organized hierarchically within schools in the same manner that individual outcomes at several points in time in the panel are organized within individuals. Outcomes to students in the same school are likely to be nonindependent for the same reason that responses by the same individual over time are nonindependent: the model will almost certainly ignore some causal variables that are stable for schools (or individuals) over pupils (or time). For example, students in the same school are exposed to some extent to a common school culture, to common facilities, have common histories, etc.

Given this close similarity there appears to be much to be gained from reformulating the hierarchical structure analysis problem in GLS terms. If we formulate the model in terms of a state space defined in terms of the lowest level units (e.g. students) and include components to variance for each higher level units, we see that in most cases the optimal statistical estimators of organizational effects will be neither the individual level model nor the between-school models discussed above. These models are appropriate only in extreme cases, namely when individual outcomes are completely uncorrelated or perfectly correlated. In all other cases the appropriate model will fall between these two and will depend on the actual levels of intercorrelation in the sample.

I suggest that the same issues arise in the analysis of the effects of aspects of university structure on student behaviors. Suppose one is interested in analyzing the effects of the nature of ties of universities with the state and the economy on the political attitudes of students. One might conduct between university studies, between school studies, individual level studies, etc. All of the reasoning just made concerning studies of achievement also appears to apply here.

The theme of this paper has been that sociologists pay too little attention to model specification other than to the problem of including the proper variables. I have tried to argue both through example and from general principles that improper state description of a process can be as damaging to inference as including the wrong variables in a model of the process. This sort of problem has high salience in research on the relations of educational organizations and other elements of social structure since the hierarchical nature of educational structures admit multiple specifications of causal processes. I have made several proposals concerning adjustment of research practice in this field. Clearly the issues demand deeper analysis.

FOOTNOTES

1. For reviews of the main features of this and the second strategy discussed below, see Blaug (1968, 1969) and Machlup (1970).

2. Kuznets (1966) emphasizes such information flows as a major determinant of national economic growth.

3. A third alternative is to conduct (pooled) within-school (more generally, within-group) analyses. Cronbach (1976) argues forcefully that while either between-group or within-group analyses may be appropriate, the total individual analysis is not. Clearly a preference for within-group analysis presumes that the boundaries around the groups are very strong (and substantively meaningful). This may be the case for analyses of instruction in certain types of primary and secondary school organizations. It is unlikely to be the case in analyses of higher educational organization. Few of the units within universities are tightly bounded. Therefore, I will not consider within-group analyses here.

REFERENCES

Bidwell, Charles E. and John D. Kasarda
1975 "School district organization and student achievement." American Sociological Review 40: 55-70.

1976 "Reply to Hannan, Freeman, Meyer, and Alexander and Griffen." American Sociological Review 41: 152-160.

Blalock, Hubert M. Jr. (ed.)
1971 Causal Models in Social Research. Chicago: Aldine.

Blaug, M. (ed.)
1968 The Economics of Education I. Baltimore: Penguin Books.

1969 The Economics of Education II. Baltimore: Penguin Books.

Coleman, James S.
1968 "Mathematical models of change." In H. M. Blalock, Jr. and A. Blalock (eds.) Methodology in Social Research. New York: McGraw-Hill.

Cronbach, Lee J.
1976 Educational Effects at the Individual and Group Levels: The Crucial Importance of Hierarchical Interpretation. Final Report to the Spencer Foundation.

Feldman, Kenneth A. and Theodore M. Newcomb (eds.)
1969 The Impacts of College on Students. San Francisco: Jossey-Bass.

Freeman, John and Michael T. Hannan
1975 "Growth and decline processes in organizations." American Sociological Review 40:

Goldberger, Arthur S.
1973 "Efficient estimation in over-identified models: an interpretive overview." In A. S. Goldberger and O. D. Duncan (eds.) Structural Equation Models in the Social Sciences. New York: Seminar Press.

Goldberger, Arthur S. and Otis Dudley Duncan (eds.)
1973 Structural Equation Models in the Social Sciences. New York: Seminar Press.

Hannan, Michael T.
1971 Aggregation and Disaggregation in Sociology. Lexington, Mass.: Heath-Lexington.

1973 Education and Economic Development: Some Baseline Models. Final Report to the Office of Education.

Hannan, Michael T. and John Freeman
1976 "The population ecology of organizations." *American Journal of Sociology* (forthcoming).

Hannan, Michael T., John Freeman and John W. Meyer
1976 "Specification of models of organizational effectiveness: comment on Bidwell and Kasarda." *American Sociological Review*. 41:136-143.

Hannan, Michael T. and Alice A. Young
1974 "Estimation in pooled-cross-section and time series models: preliminary Monte Carlo results." Proceedings of Conference on Policy Research: Methods and Implications, University of Wisconsin, Madison.

1976 "Estimation in multi-wave panel models: results on pooling cross-sections and time series." In D. Heise (ed.) *Sociological Methodology 1977*. San Francisco: Jossey-Bass (in press).

Kuznets, Simon
1966 *Modern Economic Growth*. New Haven:

Machlup, Fritz
1970 *Education and Economic Growth*. Lincoln, Neb.: University Nebraska Press.

Maddala, G.S.
1971 "The use of variance components models in pooling cross-section and time series data." *Econometrica* 39: 341-358.

Meyer, John W., Michael T. Hannan, Richard Rubinson and George Thomas
1976 "National economic in the contemporary world system, 1950-1970: social and political effects." Paper presented at conference on Polish and American Youth: Higher Education and Changes in Social Structure.

Nerlove, Marc
1971 "Further evidence on the estimation of dynamic economic relations from a time series of cross-sections." *Econometrica* 39: 359-382.

Nielsen, Francois and Michael T. Hannan
1976 "The expansion of national educational systems: tests of a population ecology model." Unpublished manuscript.

Vásquez
Associates, Ltd.
P. O. BOX 5630 MILWAUKEE, WISCONSIN 53211

Alma G. Vásquez
President

PROCEEDINGS

Conference on Methodology for Aggregating Data in Educational Research*

Stanford University
October 23-24, 1976

Organizers: Leigh Burstein
Department of Education
University of California,
Los Angeles

Michael T. Hannan
Department of Sociology
Stanford University

*Sponsored by Vasquez Associates, Ltd., Milwaukee, Wisconsin under Contract #NIE-C-74-0123 from the National Institute of Education.

TM006 473

Vásquez Associates, Ltd.
P. O. BOX 5630 MILWAUKEE, WISCONSIN 53211

Alma G. Vásquez
President

Conference on Methodology for Aggregating
Data in Educational Research

Stanford University
October 23-24, 1976

SCHEDULE

Saturday, October 23

9:45 a.m.	Conference Opening
10:00 a.m.-1:00 p.m.	Panel 1 -- Analysis of Multilevel Data Lee J. Cronbach, Research on Classrooms and Schools: Formulation of Questions, Design and Analysis Leigh Burstein and Robert L. Linn, Detecting the Effects of Education in the Analysis of Multilevel Data: The Problem of Heterogeneous Within-Group Regressions Michael T. Hannan and Alice A. Young, On Certain Similarities in the Analysis of Multi-Wave Panels and Multilevel Cross-Sections J. Ward Keesling, Components of Variance Models in Multilevel Analysis Charles N. Henderson, Jr., Comments on "Statistical Analysis of a Complex Model Involving Fixed and Random Factors, Unequal Subclass Numbers, and Higher Level Covariates"
2:00 p.m.-5:00 p.m.	Panel 2 -- Contextual Effects Duane F. Alwin, Structural Equations Models of Contextual Effects Lee J. Cronbach, Research on Classrooms and Schools: Formulation of Questions, Design and Analysis John Meyer, Research Designs Required for Contextual and Aggregate Studies of School and Schooling Effects Alice A. Young and Lutz Erbring, Unraveling Classroom Effects: A Multilevel Feedback Model of Contextual Effects Aage Sorenson, Comments on "School Effects in Growth in Academic Achievement"
8:00 p.m.	Dinner, China First Restaurant, Palo Alto

Sunday, October 24

9:00 a.m.-12:30 p.m.	Panel 3 -- Miscellaneous Topic on Units of Analysis Linda Glendening and Andrew C. Porter, Independence and Selecting the Appropriate Unit of Analysis (or the Appropriate Analytic Model) Michael T. Hannan and Alice A. Young, Small Sample Results on Estimation from Grouped Observations Carlyle E. Maw, Problems of Data Aggregation and Cross-Level Inference with Categorical Data

Vasquez Associates/NIE Conference on Methodology for Aggregating Data in Educational Research, October 23-24, 1976, Stnaford University

---

LIST OF PARTICIPANTS:

Duane F. Alwin, Dept. of Sociology, Indiana University

Leigh Burstein, Grad. Sch. of Education, UCLA

Frank Capell, Sch. of Education, Stanford University

Lee J. Cronbach, Sch. of Education, Stanford University

Lutz Erbring, Inst. for Social Research, University of Michigan

John H. Freeman, Sch. of Business Admin., University of California, Berkeley

Linda Glendening, Dept. of Health, Education & Welfare, National Inst. of Education

Arthur Goldberger, Department of Economics, University of Wisconsin

Maureen T. Hallinan, Dept. of Sociology, Stanford University

Walt Hancy, Huron Institute

Michael T. Hannan, Dept. of Sociology, Stanford University

Charles Henderson, Dept. of Human Development, Cornell University

Kjëll Harnqvist, Dept. of Psychology, University of Götenberg, Sweden

Robert L. Linn, Dept. of Psychology, University of Illinois

J. Ward Keesling, Grad. Sch. of Education, UCLA

Thomas R. Knapp, College of Education, University of Rochester

Carlyle E. Maw, Natl. Inst. of Education, Measurement & Methodology Division

Jack Merwin, College of Education, University of Minnesota

John Meyer, Dept. of Sociology, Stanford University

Francois Nielsen, Dept. of Sociology, Stanford University

Andrew Porter, College of Education, Michigan State University

Hugh Poynor, Kirschner Associates

David Rogosa, Sch. of Education, Stanford University

Aage Soreson, Dept. of Sociology, Stanford University

Nancy Tuma, Dept. of Sociology, Stanford University

Noreen Webb, Sch. of Education, Stanford University

Alice A. Young, Grad. Sch. of Ind. Adm., Carnegie-Mellon University

The following papers, which were presented at the Conference on Methodology for Aggregating Data in Educational Research at Stanford University, are also technical reports of the Consortium on Methodology for Aggregating Data in Educational Research under National Institute of Education Contract #NI#-C-74-0123, Vasquez Associates, Ltd., Milwaukee, Wisc. 53211

T.R.#

15. Webb, N., Capell, F. and Nielsen, F. "Summary of Presentation and Discussion, Conference on Methodology for Aggregating Data in Educational Research," Stanford University, October 23-24, 1976. Technical Report No. 15, October, 1976.

16. Burstein, L. and Linn, R.L. "Detecting the Effects of Education in the Analysis of Multilevel Data: The Problem of Heterogeneous Within-Group Regressions." Technical Report No. 16, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

17. Hannan, M.T. and Young, A.A. "On Certain Similarities in the Analysis of Multi-Wave Panels and Multilevel Cross-Sections." Technical Report No. 17, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

18. Keesling, J. W. "Components of Variance Models in Multilevel Analysis." Technical Report No. 18, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

19. Henderson, Charles N. Jr. "Comments on 'Statistical Analysis of a Complex Model Involving Fixed and Random Factors, Unequal Subclass Numbers, and Higher Level Covariates'" Technical Report No. 19, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

2[illegible] Alwin, D. F. "Structural Equations Models of Contextual Effects." Technical Report No. 20, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

21. Meyer, J. "Research Designs Required for Contextual and Aggregate Studies of School and Schooling Effects." Technical Report No. 21, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

22. Young, A.A. and Erbring, L. "Unraveling Classroom Effects: A Multilevel Feedback Model of Contextual Effects." Technical Report No. 22, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

23. Glendening, L. and Porter, A.C. "Independence and Selecting the Appropriate Unit of Analysis (or the Appropriate Analytic Model)." Technical Report No. 23, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

24. Hannan, M.T. and Young, A.A. "Small Sample Results on Estimation from Grouped Observations." Technical Report No. 24, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

25. *Maw, C.E. "Problems of Data Aggregation and Cross-Level Inference with Categorical Data." Technical Report No. 25, October, 1976. Also, paper presented at Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976.

*An earlier version of this paper was presented at the AERA annual meeting in San Francisco, April, 1976.

Conference on Methodology for Aggregating Data in Educational Research*

Stanford University
October 23-24, 1976

Summary of Presentations and Discussion

by

Noreen Webb, Frank Capell, and Francois Nielsen
Stanford University

---

*The Conference was organized by Vasquez Associates, Ltd., Milwaukee, Wisconsin under Contract #NIE-C-74-0123 from the National Institute of Education.

Saturday Morning

Leigh Burstein (LB) opened the conference with a description of the purpose of the sessions: to share current thinking on the problem of units of analysis. Each presentation summarized the gist of the paper, and was followed by discussion. Saturday morning's session was to be devoted to analyses of multilevel data. Saturday afternoon's session was to focus on contextual effects. LB emphasized that the two topics were closely related and that discussion should blend from one topic to another. In Sunday morning's session, miscellaneous topics on units of analysis were to be discussed.

Lee J. Cronbach (LJC). Research on Classrooms and Schools: Formulation of Questions, Design and Analysis.

LJC introduced his paper describing his concern that we have failed to get good models, that the problem is substantive (underspecification) rather than statistical. Learning is individual but takes place in groups. There is a severe limitation on the number of groups in a study. The question is one of process: learning usually takes place in groups. So, we should examine learning *within* groups.

LJC summarized the usual regression formulation.

1. Everyone has tried to explain outcome Y as a function of individual X_{pc} first, and then $\bar{Y}$ of the group as a function of group $\bar{X}_c$.

2. $Y = f(X_{pc})$ and $\bar{Y} = f(X_c)$ are redundant. We want to separate $\bar{Y} = f(\bar{X}_c)$ and $Y = f(X_{pc} - \bar{X}_c)$.

LJC used as an example a study by Bowers (Bowers, 1968) in which Bowers asked students in 99 colleges whether they disapproved of drunkenness. Bowers thinks the difference between the group regressionand individual regression (see diagram below) is evidence of a contextual effect. But LJC prefers to look at the within-group regressions.

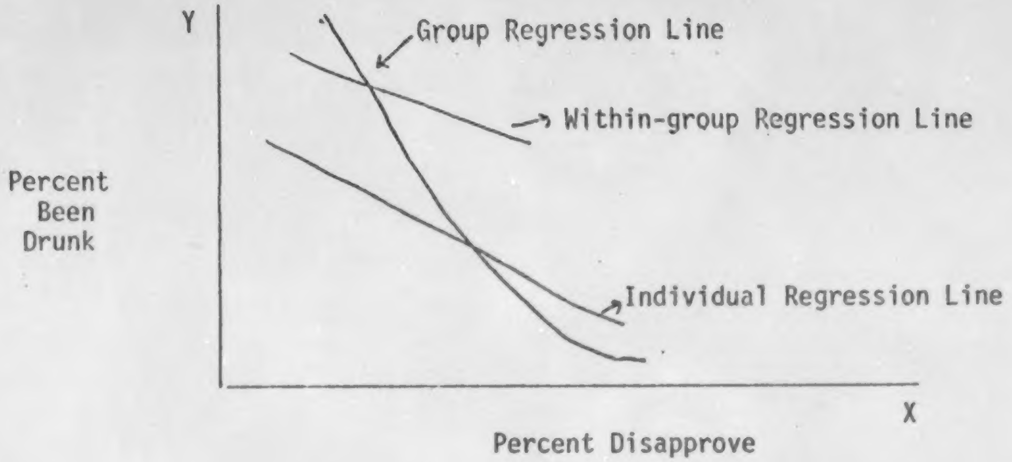

Y
Group Regression Line
Within-group Regression Line
Percent Been Drunk
Individual Regression Line
X
Percent Disapprove

Figure 1. Diagram depicting difference between the group regression and individual regression as evidence of a contextual effect.

LJC described the importance of discriminants to describe initial differences among groups: what causes people to be separated into groups? There may be a number of interesting substantive-looking effects that are artifactual. LJC closed with the question: "How can we separate 'demographic' (artifactual) and substantive effects?"

In response to the question by Carlyle Maw (CM) about whether we should find the discriminant function for grouping variables, LJC said the following. Grouping variables are part of the mathematical model -- as grouping rules. Even though in the multivariate case there are several grouping variables, in actual data we will observe group means. We should assess all prior variables of importance. LJC referred to the Feige-Watts problem: what is a sensible way to constitute aggregates if we have individual data (which may be done for security reasons, for example). LJC's concern is what happens when the groups are already constituted.

Leigh Burstein (LB) and Bob Linn (BL): Detecting the Effects of Education in the Analysis of Multilevel Data: The Problem of Heterogeneous Within-Group Regressions.

LB described several components that enter pupil outcomes:

(1) Adjusted effects for classrooms: $\bar{Y}_{.c} - \beta_b(\bar{X}_{.c} - \bar{X}_{..})$,

(2) Pooled within-class regression: β_w ,

(3) Specific within-class regression: $\beta_c - \beta_w$.

The last reflects what specific teachers do to influence different students within the classroom. It is important to understand the differences from one class to another. One classroom may exhibit compensatory effects; others may show meritocratic effects; still others may demonstrate random effects within the class. LB stressed the importance of looking at class slopes, not only class means. Even if teacher quality has a small effect on the slope, the effect in the average outcome may be dramatic. So a straight forward analysis of class means is not enough; slopes may have information in addition to means.

LB concluded with the remarks that we should look more to internal properties of within-group process; and that, even though inelegant, pooling may be in the end, the only way to obtain the desired information.

Following LB's presentation, Art Goldberger (AG) suggested that differences in specific within-class slopes may be attributed to pre-existing differences rather than teacher effects. LB agreed.

Mike Hannan (MH) and Alice Young (AY). On Certain Similarities in the Analysis of Multi-Wave Panels and Multi-Level Cross-Sections.

MH opened his presentation with the comment that the question of whether to use between or within-group regression can cycle endlessly. He suggested looking at the problem from a different perspective; that of pooled cross-sectional and time series models. We should get reasonable predictions for within and between effects from a well-specified model.

MH described the models, asking whether there is an issue to decide between the different models. The choice is the same in panel analysis; we can average in several ways: the pooled total is an average over time and individuals; the between-individual is an average over time; and the between-time is an average over individuals. No one will accept the pooled total because responses of individuals change over time.

MH described the use of GLS in this problem. Few people realize, he said, that we cannot usually apply OLS to pooled data, because of the redundancy in the data. The GLS estimate combines between-individual and within-individual variation in an optimal way. The same problem arises in pupil and classroom analysis: pupils are not independent within a class. The optimal model is not the total model. The optimal model may be GLS to weight individual and group variation.

The issue is more complicated, he explained, when the model is improperly specified. Some omitted variables may vary between-classes but not within-class, and others may vary within but not between. The bias can be different in the total vs. between vs. within. The bias also depends on the particular model. If one can identify left-out variables, then GLS can be easily adapted to the problem.

In an interchange between Carlyle Maw and MH, MH explained that the variance-covariance matrix is estimated from the sample data. In GLS, he explained, we have to know the true variance-covariance matrix, so we are really talking about the general model. Francois Nielsen (FN) added that it is possible to input substantive assumptions using constrained estimation.

J. Ward Keesling (WK). Components of Variance Models in Multilevel Analysis.

WK introduced his paper with several comments about Schmidt's (1969) dissertation model and the distinction between random and fixed effects. If classes are drawn randomly, it is possible to look at components of variance models.

WK explained that the between-class effect still contains within-class variability: $E(MS_B) = \sigma^2 + n\sigma_a^2$ where σ^2 is the within-group variability. In the Schmidt model, where $\Sigma_y = \Sigma_a + \Sigma$, a may not be positive definite. A separate parameterization of each matrix is suggested. Two possible specifications are $\Sigma_a = \Lambda_a \Phi_a \Lambda'_a + \Psi_a$ and $\Sigma_a = \Lambda \Phi \Lambda' + \Psi$, depending on the level of analysis. It is possible to check the fit of the model under different specifications in within-group or between-group space. WK described an example of a just-identified model, in which the relationship between posttest and instructional time was different in the two specifications above.

In the interchange between Duane Alwin (DA), WK, Lee J. Cronbach (LJC), and Charles Henderson, Jr. (CH), DA asked whether the disturbance matrix is non-null in general. WK maintained that it depends on empirical data. LJC discussed the larger issue: what kinds of events can reasonably be modeled? Can we reproduce the situation? A random model is not applicable to classrooms; when we randomly replace members of a class, we have a new organism. The critical issue becomes: when can we talk about random groupings? When is the pupil or the class the unit randomly chosen?

CH then described how it is not possible to have students fixed within random classrooms. If the design has repeated measures on pupils, then we can consider the pupil as random. LJC countered with the example of research on marriages, in which families are random and husband and wife are fixed. If one member of a pair is drawn, then that member is considered fixed. Henderson concluded the interchange noting that the model must be fully specified, in which case some factors may be fixed.

Charles Henderson, Jr. (CH). Comments on "Statistical Analysis of a Complex Model Involving Fixed and Random Factors, Unequal Subclass Numbers, and Higher Level Covariates".

The following are the main points of CH's presentation of his three papers:

(1) In general, there are no appropriate test distributions. Most tests are biased. If one ignores a random factor, it is possible to obtain unbiased estimates of fixed effects, but the tests are biased. If classroom is considered as a fixed effect, the estimators may be biased.

(2) There is danger in the current analysis of variance programs. If the higher-level covariate is a fixed factor, it cannot be estimated. The only solution is to treat higher-level covariates as random factors.

(3) The choice of the unit of analysis is the wrong emphasis. Instead, we should fully specify the model, with all fixed and random factors and then choose the best method of estimation.

(4) CH described an ecological model concerning socializing agencies, asking the question: are there ethnic differences in this process between cities in Israel? The individual in the model is a level above the base line. Some interactions among random factors are eliminated to make the model estimatable.

The following points summarize the discussion between CH, Tom Knapp (TK), Lee J. Cronbach (LJC), Mike Hannan (MH), Carlyle Maw (CM), and Lutz Erbring (LE).

(1) CH described that if the error components are known, the best method of testing the model is GLS, which gives unbiased estimates. It is possible to estimate effects as if all factors are fixed.

Higher-level _and_ lower-level covariates must be included in the equation. If classroom is the unit of analysis, then classroom should also be in the model. Then it is possible to test whether the within- and between-class regression slopes are different from zero, and whether the within-class regression differs from the between-class regression.

(2) TK raised the question of whether the total regression is ever interesting. LJC explained that if we are interested in causal processes, different things happen between- and within-classes. If we are interested in a gross summary, the total regression may suffice.

(3) LJC described how the hypothesis of no group effects is not plausible in social science research. When the data allow looking at separately the within- and between-group effects, it is worth doing.

(4) MH raised the issue of the group boundary. Sometimes the class is too large a unit; there may be subgroups within the classroom. CM described how, in educational sociology, the units may not be modifiable -- as in political units of school, district, class, etc. MH countered that we should not rectify the unit of analysis, that it may be possible to partition the classroom effect into between-clique regressions.

(5) CH emphasized again that all possible factors should be included in the model.

(6) LE described how some higher-level factors represent ignorance about the individual-level processes. If the model were better specified, the effect of the class as a special factor may not be needed. Two ways to remove specification bias may be to have micro-theory or to include potent higher-level covariates, and consequently not need to use proxy variables.

Saturday Afternoon

Duane Alwin (DA). Structural Equation Models of Contextual Effects.

In introducing his topic, DA mentioned that the present paper is a continuation of work reported in Hauser, Sewell and Alwin.

He listed some general concerns:

1) Sociologists have typically ignored classrooms in favor of individuals and schools. He suggested some reasons:
   a) When measuring teachers and pupils, classrooms are not as important a source of additional variation.
   b) It is more typical of sociologists to focus on later school life (e.g., high school or college) when the intact classroom is a less relevant concept.
   c) Since classroom variables are not under direct experimental control, we often don't know what the treatments are.
2) Sociologists typically use cross-sectional data. Occasionally there will be a longitudinal or follow-up design, but rarely, if ever, do sociologists deal with change, with lagged dependent variables. The exception is in Sorenson and Hallinan's work. The latter is a valuable perspective that traditionally sociologists have ignored.
3) The third point has to do with selection problems. We try to explicitly represent selection-type variables -- the X's in Cronbach's notation -- in our models. But, we really have little idea about processes that might be produced by selection on the dependent variable. That is, the selection variables are usually independent variables.
4) It seems that the "action" is with the within-group variation, with the between-group variation regarded as a source of contamination.
5) Policy relevant variables are more likely to be at the between-group level.

6) Semantic/terminological problems persist: contextual vs. compositional vs. school vs. schooling effects.

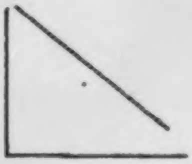
Other introductory comments:

7) Problems of reliability are probably best addressed in the framework of the work of Jöreskog.

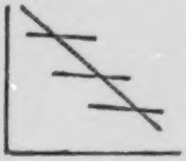
8) DA defines the contextual effect as the difference between the between and the within slopes.

9) Substantively, DA finds little evidence for differences between within and between slopes; there are few real contextual effects.

Lee J. Cronbach (LJC) expressed concern about the percentage of variance type arguments used, for example, by Hauser.



①



②

For example, even though ② may be the true model, ① might be thought adequate using % variance criterion.

DA responded that in his view, % variance is not meant as a rigid criterion, and asked what other criterion might be proposed.

LJC answered that he would be more interested in looking at what goes on in the upper and lower quartiles of the X-distribution, albeit at the cost of parsimony.

The discussion shifted to a consideration of specific equations in the Alwin paper:

DA said he used eq. (15) for empirical estimation;

Charles Henderson (CH), agreed that (15) was the correct one for obtaining unbiased estimates in higher-level regressions, adding that eq. (2) would produce biased estimates.

Nancy Tuma (NT) commented that eqs. (2) and (15) seemed to represent different substantive models.

DA said eq. (28) is a variant of eq. (15) used in the paper, the term of interest being $b_B - b_W$ which he calls the contextual effect.

LJC likened this contextual effect to what he refers to as a demographic effect.

DA:

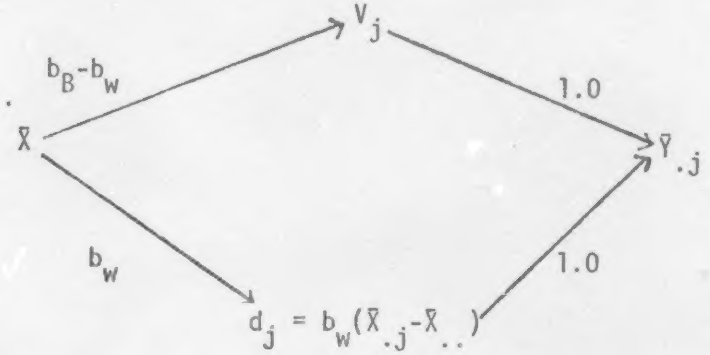


Diagram represents a separating of the effects of $\bar{X}$ on $\bar{Y}$ into the compositional effect (bottom path) and the contextual effect (upper path).

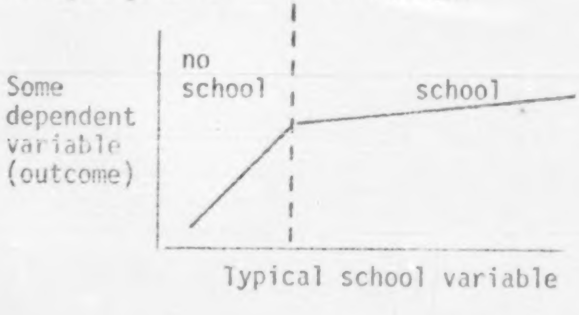
LJC expressed concern for the possibility of incomplete adjustments present in such models.

Mike Hannan (MH) interpreted LJC to say: if students are grouped in a systematic way, contextual effects may be formed even if in reality there are none.

MH added that if you can put the grouping variables on the right-hand side of the equation, the problem is taken care of.

John Meyer (JM). Research Designs Required for Contextual and Aggregate Studies of School and Schooling Effects.

JM began with a quick review of substantive findings on school effects, using the following figure as an illustration:



JM said that little variation in outcome is observed when students are compared among themselves, but large differences are found in comparisons of students and non-students.

He then outlined three theoretical formulations which have been suggested as explanations for the absence of school effects:

1) Teaching quality is the critical variable, but teachers are more or less randomly assigned to schools, equalizing aggregate teacher quality across schools.
2) Effectiveness of schools is due to "time on task" type variables which do not vary widely from school to school (cites David Wiley as proponent of this view).
3) The relevant property of schools is simply that they are schools (licensing agencies), and people adapt their thinking to this fact (JM the proponent of this view).

JM: Research suggestion following from the 3rd view would be to study outcomes associated with different types of curricula (e.g., college prep. vs. non-college prep.), where curriculum is formally regarded as a labelling variable.

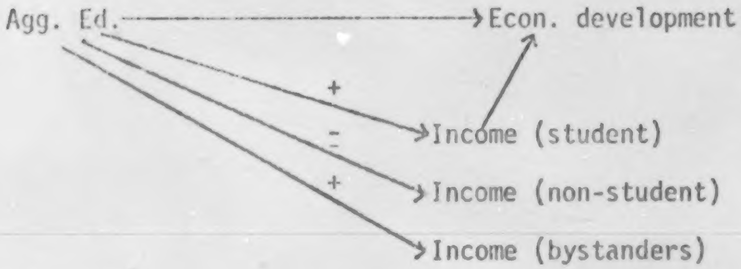
DA commented that, for example in the German and French educational systems, type of school is confounded with curriculum, and also enormous selectivity.

Arthur Goldberger (AG) adds self-selection is a confounding factor.

DA then asked how context and curriculum may be separated.

JM responded: hold school constant and vary credentialing system of education in a country -- need variation in the independent variable before effects can be discussed.

JM then went on to talk about the issue of aggregation, beginning with the statement that for the classroom unit there seems to be no theory which operates without appeal to individual-level mechanisms. He presented the following diagram as an illustration of a possible aggregate model.



Lutz Erbring (LE) commented that the issue is whether the variance in schooling is related to the variance in income.

Aage Sorenson (AS) suggested that education may be a purely allocative device and that structure is determined by different factors.

JM agreed.

Arthur Goldberger (AG) reported on a paper he heard at a recent conference relevant to the issues raised in JM's presentation. The paper dealt with group vs. individual economic returns of education, and suggested that for individuals, the rate of return was low, but for larger aggregates, the social benefits of education are obvious.

JM closed, stressing the need for comparative research, suggesting that the important processes in such research may have little to do with individual level variables.

Lutz Erbring (LE), Alice Young (AY) and Lutz Erbring. *Unraveling Classroom Effects: A Multilevel Feedback Model of Contextual Effects.*

LE prefaced his comments by expressing a desire to attempt to revive the discussion of contextual effects after it had been killed by the aggregation problem.

LE: Once it is recognized that many contextual effects are spurious, there may still remain some intrinsic group properties. He prefers not to define context in statistical terms but assumes there are real group processes. He suggested starting by specifying the process rather than the effect. Such processes may be described by models incorporating edogeneous feedback: if performance of one student affects performance of others, then notions such as dyadic linkages, mutual awareness, face-to-face interaction become relevant. Exogenous variables feed back into the system through classroom interaction. The models required are simultaneous equation systems with endogenous variables on both sides of the equations. The simultaneous nature of the equations, even using reduced forms, poses difficult estimation problems: in practice, special methods are required (cf. Lebanon, et al., 1976). The model can be extended to multi-classroom situations by assuming the contingency structure to be more or less equal among units with classes, and zero between classes. In certain instances, it may be appropriate to summarize endogeneous feedback by including $\bar{y}$ on the right hand side of the equation.

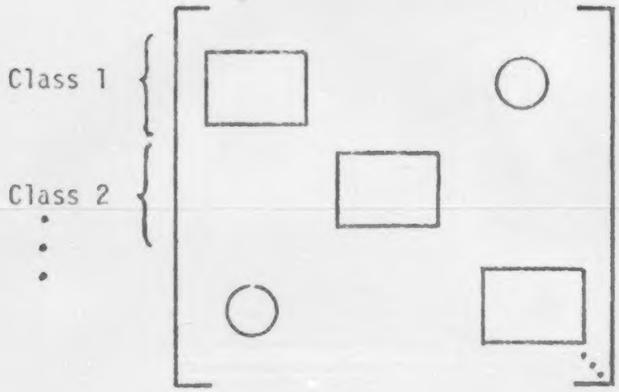
If feedback is ignored, as in the usual analysis, the correlation of the omitted factor with the exogeneous variables will result in biased estimation: for example, including friends' aspirations in attainment model -- "statistical" contextual effects will persist even with good specification.

Aage Sorenson (AS) commented that the model seems to require sociometric type data, which are known to be quite unreliable, and suggested alternatively specifying the model over time, using differential equations.

LE responded that it is also possible to make simplifying assumptions (e.g., coefficients equal within classes).

Art Goldberger (AG) asked for clarification of the special case with zero between class and equal within class.

LE: The model is $y = \alpha Wy + X\beta + \mu$. The zero between, equal within case can be pictured as follows:



with $W = I \otimes \ell\ell'/n \Rightarrow y = \alpha\bar{y} + x\beta + \mu$ where $\bar{y}$ is the v~~i~~ector of class means.

Duane Alwin (DA) asked LE to state again what type of feedback processes could lead to the contextual effects sometimes found.

LE: The problem is the bias in the model $y = x\beta + \mu$ when the true model is $y = \alpha Wy + x\beta + \mu$.

For the underspecified model:

$$E(\hat{\beta}) = \beta + E[(X'X)^{-1}X'Wy],$$

which will be unbiased only when $W = 0$. The bias is the difference in the between and within slopes.

Aage Sorenson (AS). Sorenson-Hallinan, Comments on "School Effects in Growth in Academic Achievement".

AS first commented that his approach parallels that of Erbring & Young, but differs in important ways.

The problem is how to incorporate contextual effects in a model, then how to estimate them. The reasoning is as follows: conceptualize learning process as it takes place in schools, focussing on the mechanisms by which it occurs. He identifies three factors in learning: ability, opportunity and effort.

The model: $V(t)$ -- amount taught at time t

V^* -- amount to be taught

S -- ability and effort

$y(t)$ -- amount learned at time t

$\frac{dy(t)}{dt}$ -- learning rate (instantaneous rate of change in amount learned)

$$dy(t) = Sdv(t)$$

$$\frac{dV(t)}{dt} = b[V^* - V(t)], \text{ solving gives:}$$

$$V(t) = V^o(1 - e^{-\frac{1}{V^*}t})$$

$$b = \frac{1}{V^*}, \text{ assuming } y(0) = 0$$

b can be taken as the opportunity for learning.

$$y(t) = \frac{s}{b}(e^{bt} - 1)$$

$$\frac{dy(t)}{dt} = s + sy(t)$$

$$y(t) = \frac{s}{b}(e^{bt} - 1) + e^{bt}y(0)$$

let $s = a + \Sigma(a_i - x_i)$ that is, s is a function of exogenous variables,

then $y(t) = y(0)e^{bt} + \frac{a_o}{b}(e^{bt} - 1) + \sum_i \frac{a_i}{b} X_i$

$$y(t) = y(o)e^{bt} + \frac{a_o}{b} - \sum_i \frac{a_i}{b}(e^{bt} - 1)X_i$$

$$y(e) = -\frac{a_o}{b} - {}_i\frac{a_i}{b} X_i$$

AS: (1) We can obtain a cross-sectional model by letting N → ∞ in terms of dynamic parameters.

(2) The implication of the model is that good schools increase educational inequality.

(3) There seems to be a general lack of good data for estimating this type of model.

Duane Alwin (DA) commented that there may be a lot of studies of time-lagged achievement.

AS responded, that educational psychologists have good over-time data but poor individual data while sociologists have poor time data and a lot of individual data (sociometric?).

Maureen Hallinan (MAH) commented that although the assumptions used in deriving the equations are sometimes ad hoc, the model could be refined given more attention to the learning process.

Carlyle Maw (CM) commented that he felt some tension between the formulating of more elaborate models, and the lack of data to test them with; he sees the morning and afternoon sessions at cross-purposes.

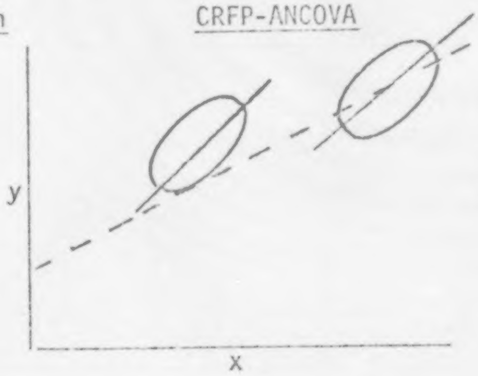
Lutz Erbring (LE): The mathematical models give rise to estimates of phenomena that are precisely the ones discussed in the morning session.

AS: The differential equation models can be estimated from linear regression coefficients.

Mike Hannan (MH): In both cases we finally get to linear models, but they are derived theoretically.

Lee J. Cronbach

CRFP-ANCOVA



Pooled regression slope -- will frequently lead to underestimation.

Two important concepts:

1. The ideal covariate (ξ^*) -- all information about the individual that predicts y within treatments.

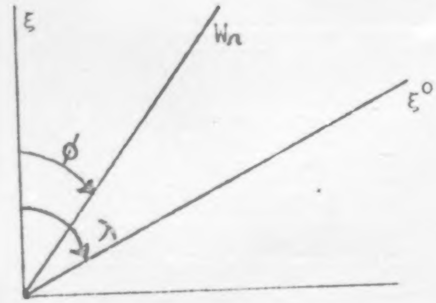
2. The discriminant (ξ^0) -- all information that determines what treatment an individual will get (e.g., self-selection, etc.)

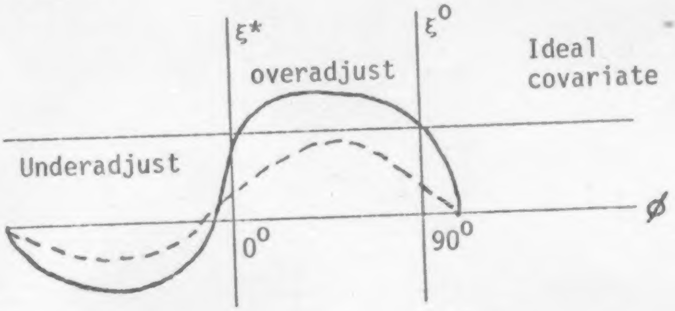
$$y = f(\xi^*);\ T = f(\xi^0)$$

$\Omega = a\xi^* + b\xi^0$ the set of all linear combinations of ideal covariate and complete discriminant.

$W = W_\Omega \in \Omega + u$ -- an observed covariate

u attenuates the adjustment





-----: $\geq$ 50% noise (i.e., u) produces underadjustment everywhere.

Sunday A.M.

Linda K. Glendening (LG) and Andrew C. Porter (AP), Independence and selecting the appropriate units of analysis (or the appropriate analytical model).

The oral presentation by (LG) was a verbatim reading of the conference paper (see attached).

The key points of the presentation:

Linda Glendening (LG) stated an interest in the class of designs which attempt to test for effects of interventions which are hoped to affect individual performance, but which are generally delivered to groups of individuals, e.g., classrooms. More specifically, she and Porter attempted to investigate the effect of violations of the assumptions of independence on tests of the hypothesis of no treatment effects on a dependent variable; or alternatively, the effect of testing that same hypothesis with an incorrectly specified model.

The study reported in the presentation involved three stages:

(i) derivation of an operational definition of independence of responses;

(ii) use of the derived definition to study the effect of violating the assumption of independence within a ANOVA context;

(iii) examination of the questionable procedures of doing a preliminary test to select a proper unit of analysis to use in a subsequent test of treatment effects.

LG proposed that independence be operationally defined as that condition where the expected mean square between groups equals that within groups. Two types of dependence were also defined: positive (expected mean square between groups greater than within groups) and negative (expected mean square within groups larger than the between groups).

LG reported on a simulation of the effects of violating the assumption of independence within the context of a balanced, two-level hierarchically-nested design, having subjects nested within groups and groups nested within treatments. In all three ANOVA models were compared: the "never-pool" model with the group or classroom as the unit of analysis, the "always pool" model with the student as the unit of analysis, and the "sometimes pool" or conditional model where a preliminary test of independence (equality of expected mean squares between and within classes) is followed by a choice of unit (student or classroom) for testing treatment effects.

The results of the simulation indicate that all three models were appropriate given independence with the "always pool" model being the most powerful. Given positive dependence, the proper unit was the group ("never pool" model) while the "always pool" and "sometimes pool" models yielded spuriously small error terms and therefore too liberal tests of treatment effects. Given negative dependence, the correct unit was also the group,though in this case the "always pool" and "sometimes pool" models cause conservative F tests and spuriously reduced the empirical power of the study.

LG concluded that the preliminary test is not a viable alternative to the a priori selection of the class (dependence) or student (independence) as the unit of analysis but acknowledged the complications of obtaining prior knowledge about independence of responses.

LG also briefly addressed the issue of ANCOVA and dependence. She stated that the question of interest was "can ANCOVA correct for dependence in student within-class data, allowing student to be used as the unit of analysis?" LG points out that in ANCOVA, it is important to know the source of the dependence in the posttest variable (Y),(e.g., the covariable (X) or the treatment). Three models paralleling those in the ANCOVA situation were described. The use of the analogue to the "always pool" model (adjustment by the pooled within-treatment regression with students as units) was again judged to be inappropriate, while the choice between the pooled within-treatment regression with classes as units and the pooled within-class regression with pupils as units depends on the relative magnitudes of the variance of the dependent variable and its correlation with the covariate under the two models (assuming that the covariate actually results in the removal of variation of the dependent variable among classes within treatments).

actually results in the removal of variation of the dependent variable among classes within treatments).

Art Goldberger (AG) raised the issue of negative dependence. He was mystified by this notion because he thought the intraclass correlation was a variance ratio.

David Rogosa (DR) explained that the *square* of intraclass correlation is the variance ratio and that the lower bound of the intraclass correlation is -1/(n-1).

AG and DR, on the one hand, and Andy Porter (AP), on the other, disagreed on whether one could generate population values of the intraclass correlation lower than -1/(n-1).

AP explained his position that one can create a situation in which variance within classrooms is much greater than variance across classroom means.

AG and DR countered that zero variation among class means is possible and in that case, the pooled within-group variance will be the same as the total variance.

Everyone agreed that it is possible to have more within-group variation than between-group variation, but AG preferred to call this *no* dependence rather than *negative* dependence.

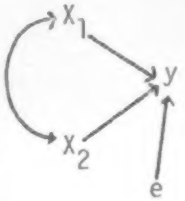
AP wrapped up the discussion with the point that, if the grouping variable is ignored when the variation within levels of the grouping variable is much greater than the variation of group means, then a conservative test of the treatment effect will result.

Alice Young (AY) Hannan & Young -- Small sample results on estimation from grouped observations

AY and Michael Hannan (MH) simulated models of the type:

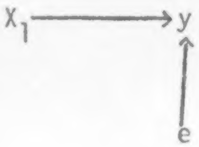
True Models

(a)



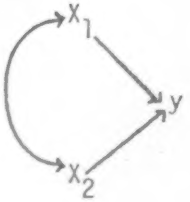
,

(b)



Models Estimated

(c)



, (d) $X_1 \longrightarrow y$, (e) $X_1' \longrightarrow y$

where $X_1' = X_1 + e_X$

They used 100 data sets of 500 individuals each, estimating first the individual level regressions, then estimating regressions on grouped data. They considered four types of groupings: by the value of X_1, by X_2, by y, and random grouping. They also used different values of R^2, $r_{X_1X_2}$ and e_X.

Bias introduced by grouping:

(1) For the true model (a), estimating (c) will be subject to bias only when grouping is by y, the dependent variable. Estimating (d) for (a) (i.e., a misspecified estimation model) will result in amplified

specification error in the presence of grouping, when grouping is by y or the omitted X variable.

(2) In the case of measurement error, grouping does not always make the bias worse: grouping by X_1' yields the same estimate as for the ungrouped data.

Lee J. Cronbach (LJC) asked about the relationship between the patterns of bias (for different values of R^2, $r_{X_1X_2}$, etc.) AY and MH obtained and the curve presented in Chapter III of his "units" paper.

Duane Alwin (DA) asked about frong pond effects where the individual and macro level coefficients may be larger or smaller than one another.

AY responded that the signs of the various covariances may be the explanation.

Art Goldberger (AG) commented that there is some danger in using simulations. Refering to the tables in their handout, he made the point that even with 100 samples, the simulations had not stabilized. He added that he felt their results could also be obtained analytically.

AY asked to what extent the analytic results depended on normal distribution assumptions.

LJC responded that distributional properties do not enter into the problem at this stage.

AG: Normal assumptions are OK. Grouping is by one variable, therefore, given a trivariate normal distribution and correlations among the variable, the grouping is implicitly by the other variable. The group means are collinear, and as cell sizes increase, the r's will go up.

Carlyle Maw (CM). Problems of Data Aggregation and Cross-Level Inference with Categorical Data

CM introduced his paper with some comments about the increasing amount of qualitative data. In some of this data, the problem of aggregation is already solved. CM used Goodman's approach to cross-level inference based on the ratio (p) of the cell frequency (unobserved) and the marginal frequency (observed). The issue is this: is there a relationship between the ratio p and the regression coefficients across groupings? The regression coefficients are stable across levels.

CM compared Goodman's conditions for cross-level inferences with those of regression. He presented two examples: grouping by X, and grouping idiosyncratically. When grouping by X, he obtained accurate estimates of the internal parameters (probability of cell entries) allowing some cross-level inferences to be made. When grouping idiosyncratically, he obtained terrible estimates of cell entries, and no cross-cell inferences could be made.

CM closed with the suggestion that Goodman's conditions, underlying the regression technique for cross-level inference, may be too strong for real data.

In the ensuing exchange between Aage Sorenson (AS), CM, Nancy Tuma (NT), and Leigh Burstein (LB), the following points were made:

1) The issue between log linear and linear approaches to contingency tables should be settled first because they lead to very different conclusions.

2) CM's discussion was independent of any decision on the choice of multiplicative or linear relationships. Rather, he presented a way of predicting conditional probability using a linear model.

3) Goodman says that if we want to estimate the ratio p, it has to be identical across subtables.

DRAFT

The Effects of Education in the Analysis of Multilevel Data: The Problem of Heterogeneous Within-class Regressions*

Leigh Burstein
University of California, Los Angeles

Robert L. Linn
University of Illinois

---

*Paper prepared for presentation at a Conference on Methodology for Aggregating Data in Educational Research, sponsored by a grant from the National Institute of Education (NIE-C-74-0123) to Vasquez Associates, Ltd., co-organized by Leigh Burstein and Michael T. Hannan, Stanford University, Stanford California, October 23-24, 1976.

DETECTING THE EFFECTS OF EDUCATION IN THE ANALYSIS OF MULTILEVEL DATA: THE PROBLEM OF HETEROGENEOUS WITHIN-CLASS REGRESSIONS[1]

It has become increasingly evident that efforts to identify the effects of education have suffered from a failure to understand complications caused by the multilevel character of educational data. Schools are aggregates of their teachers, classrooms and pupils, and classrooms are aggregates of the persons and processes within them. This being the case, the effects of education exist in one form or another both between and within the units at each level of the educational hierarchy. Yet the majority of studies of educational effects has restricted attention to either overall between-student, between-class, or between-school analyses.

Cronbach (1976) convincingly argues that the majority of studies of educational effects carried out thus far conceal more than they reveal, and that "the established methods have generated false conclusions in many studies" (p. 1). His concern is foreshadowed in the exchange among Wiley, Bloom, and Glaser as recorded in Wittrock and Wiley (1970), and by Haney's (1974) review of the units of analysis problems encountered in the evaluation of Project Follow Through.

In this paper, we attempt to focus on a concrete manifestation of the complexities alluded to above. We are interested in determining the appropriate analysis procedures when the data consist of pupils sampled randomly from schools and the primary questions concern the effects of teachers on pupil performance with only secondary interest in the effects of more global school characteristics. Research on classrooms such as that carried out by the Far West Laboratory For Educational Research in their Beginning Teacher Evaluation (BTES) Study (Berliner, 1976) attempts to answer this type of question.

The analytical methods to be considered are the between-group and pooled within-group analyses as suggested by Cronbach (1976; Cronbach and Webb, 1975), a hierarchial two-stage analysis proposed by Keesling and Wiley (1974) and a strategy which treats within-group slopes and intercepts as the outcomes in a between-group analysis[2]. All three methods are multilevel in that they require the analysis of data in at least two stages for at least two levels of units. If one accepts the basic premise that educational effects occur both within and among units, then multilevel analytical methods are called for. For the most part, we treat only those problems associated with examination of between-class (teachers) and within-class pehnomena, a two-level analysis problem.

ISSUES IN THE CHOICE OF UNITS OF ANALYSIS

Traditionally, a variety of competing points have been cited as justification for the choice of either pupils or groups (classrooms, schools, etc.) as the appropriate units of analysis in studies of educational effects. Both conceptual and statistical arguments have been voiced in favor of either level. A few of the key arguments for pupils or groups as units of analysis and selected relevant reference are stated below:

(1) (Conceptual) -- In education the phenomena we wish to investigate are pupil outcomes. More specifically, we want to determine the effects of the school (class, teacher) resources an individual pupil receives, his background and the influence of his community setting and peers on his educational outcomes (Averch et al., 1972; Burstein, 1975, 1976b; Burstein and Knapp, 1975; Burstein and Smith, 1976). Therefore, pupils are the units at which questions must finally be answered.

(2) (Conceptual) -- Pupils react as individuals and the effects on them should be the focus of educational evaluation (Bloom in Wittrock and Wiley, 1970, p. 271ff.)

(3) (Conceptual) -- Effects in the classrooms are an aggregation of effects of environmental arrangements on individuals (Glaser in Wittrock and Wiley, 1970, pp. 271ff).

(4) (Conceptual and Statistical) -- Theoretical arguments concerning the effects of educational structure on pupil outcomes are formulated at the pupil level. To analyze the data at the group level is to enhance the likelihood of specification bias and aggregation bias (Hannan, Freeman and Meyer, 1976).

(5) (Conceptual) -- The appropriate unit of study in educational evaluation is the collective -- class or school -- rather than the individual. The effects of a treatment on the classroom are fundamentally different from the effects of the treatment on the individuals within the classroom (Wiley in Wittrock and Wiley, 1970, pp. 271ff).

(6) (Statistical) -- The sampling unit determines the unit of analysis If classrooms are the sampling units, then classrooms are the unit of analysis (Cline _et al._, 1974; Cronbach, 1976; Ray, 1972).

(7) (Conceptual and Statistical) -- The unit of treatment defines the level of analysis. If treatments are administered to intact classrooms, classrooms are the units (same as in (6) plus Glass and Stanley, 1970; Peckham _et al._, 1969).

(8) (Statistical) -- Pupil performances within classrooms are generally correlated with each other. These dependencies dictate the choice of between-class analyses (Glass and Stanley, 1970; Glendening and Porter, 1974; Glendening, 1976).

(9) (Conceptual) -- Characteristics of the teacher take on the same value for every pupil in a particular classroom. An analysis at the pupil level overemphasizes the amount of information one has about class level variables (Keesling and Wiley, 1974).

(10) (Conceptual) -- The same manifest variables answer different questions at different levels of analyses. For example, a pupil's perception of classroom climate has a distinctly different meaning than the class aggregate perception of classroom climate. The sex of a student's teacher characterizes a social psychological effect, while the average teacher sex in a school characterizes an organizational process. Thus, the research foci should determine the appropriate unit of analysis (Burstein, 1976b; Cronbach, 1976; Scheuch, 1966).

(11) (Conceptual and Statistical) -- Dependence among observations within classrooms is a matter of degree rather than existence. Furthermore, the analysis of class means can mask between-class differences in the within-class distributions of outcomes and differences in the within-class regressions of outcomes on inputs. Thus, the use of class means as the only unit of analysis is inappropriate (Brown and Saks, 1975; Burstein, 1976; Lohnes, 1972).

(12) (Conceptual and Statistical) -- Overall between-student analyses are weighted averages of between-class and pooled within class analyses and are thus rarely advisable in educational contexts (Cronbach, 1976).

The arguments cited above are compelling and virtually unresolvable if a choice of either pupil or class as the only unit of analysis is required. The key to a potentially viable solution is contained in points (11) and (12). The multilevel character of educational data warrants analysis models suited to the identification of educational effects at and within each level of the educational system. The remaining discussion focuses on such analysis models and on one specific complication -- heterogeneous within-class slopes -- that may be encountered in their use.

DECOMPOSING PUPIL OUTCOMES AND EDUCATIONAL EFFECTS

Once the existence of group membership is acknowledged, any measure that varies over persons can be decomposed into its between-group and within-group components. The outcome, Y_{cp} of pupil p in classroom c can be decomposed into:

$$\underbrace{Y_{cp}}_{\substack{\text{INDIVIDUAL}\\\text{OUTCOME}}} = \underbrace{\bar{Y}_{..}}_{\substack{\text{GRAND}\\\text{MEAN}}} + \underbrace{(\bar{Y}_{c.} - \bar{Y}_{..})}_{\substack{\text{BETWEEN-CLASS}\\\text{EFFECT}}} + \underbrace{(Y_{cp} - \bar{Y}_{c.})}_{\substack{\text{WITHIN-CLASS}\\\text{EFFECT FOR}\\\text{PERSON p}}}$$

The effects of a variable X_{pc} on Y_{pc} as determined from a regression-based analysis can also be decomposed into between-class and within-class components. Cronbach (1976, pp. 3.1-3.11) presents this overall decomposition in the following manner. (We use a slightly different notation and nomenclature):

$$\begin{aligned}
Y_{cp} - \bar{Y}_{..} = {} & \beta_b(\bar{X}_{c.} - \bar{X}_{..}) && \text{BETWEEN-CLASSES PREDICTED}\\
& + \bar{Y}_{c.} - \beta_b(\bar{X}_{c.} - \bar{X}_{..}) && \text{ADJUSTED CLASS EFFECTS -- class deviation from the between-group regressions}\\
& + \beta_w(\bar{X}_{cp} - \bar{X}_{c.}) && \text{COMMON WITHIN-CLASS PREDICTED}\\
& + (\beta_c - \beta_w)(\bar{X}_{cp} - \bar{X}_{c.}) && \text{SPECIFIC WITHIN-CLASS PREDICTED}\\
& + \varepsilon_{cp} && \text{PERSON RESIDUAL}
\end{aligned}$$

The substantive description provided by Cronbach (1976, p. 3.4-3.11.) of the components of the decomposition deserves a careful reading but it is to lengthy to reproduce here. The key elements relevant to the present arguments are the adjusted class effects, the common within-class regression and the specific within-class regressions.

The adjusted class effects $[\bar{Y}_{c.} - \beta_b(\bar{X}_{c.} - \bar{X}_{..})]$ reflect the effect of the specific teacher/class on the mean outcome of its pupils after controlling for the effects of mean inputs on mean outcomes. As Cronbach (p. 3.5) points out, β_b reflects the tendency of groups with higher mean inputs to do better than others on the outcome measure. There is general agreement that large adjusted classroom effects are educationally desirable. Whatever analysis strategy is employed should be directed toward providing accurate estimates of this teacher effect on mean class outcome and, hopefully, should identify those teachers who achieve large effects. The BTES study mentioned earlier must generate evidence regarding this type of educational effect to fulfill its committment to provide policy relevant information about the effects of teacher skills on pupil preformance to the state and federal agencies that commissioned the study.

The common or pooled within-class regression $[\beta_w(X_{cp} - \bar{X}_{c.})]$ reflects the tendency of students above the class average on input to do better or worse on the outcome than the rest of the class. The usual educational finding is that classes tend to be strongly meritocratic (high positive β_w) i.e., students with higher inputs gain more from a given increment of instruction. An alternative egalitarian model would have teachers bringing up pupils with lower inputs at a faster rate than they increase the performances of higher ability pupils. If teachers behaved in an egalitarian fashion, β_w would still be positive, most likely, but flatter than in meritocratic classes. The comparison of the slopes of classes E and G in Figure 1 illustrates a typical egalitarian-meritocratic distinction.

Estimation of the common within-class regression is of value in the BTES study. β_w provides an indication of the overall redistributive properties of the classroom instruction encountered in the California school population represented in the study. If policy makers and teacher educators are intent on raising the relative performance of low ability and low SES youth, through better teaching methods and more skilled teachers, β_w

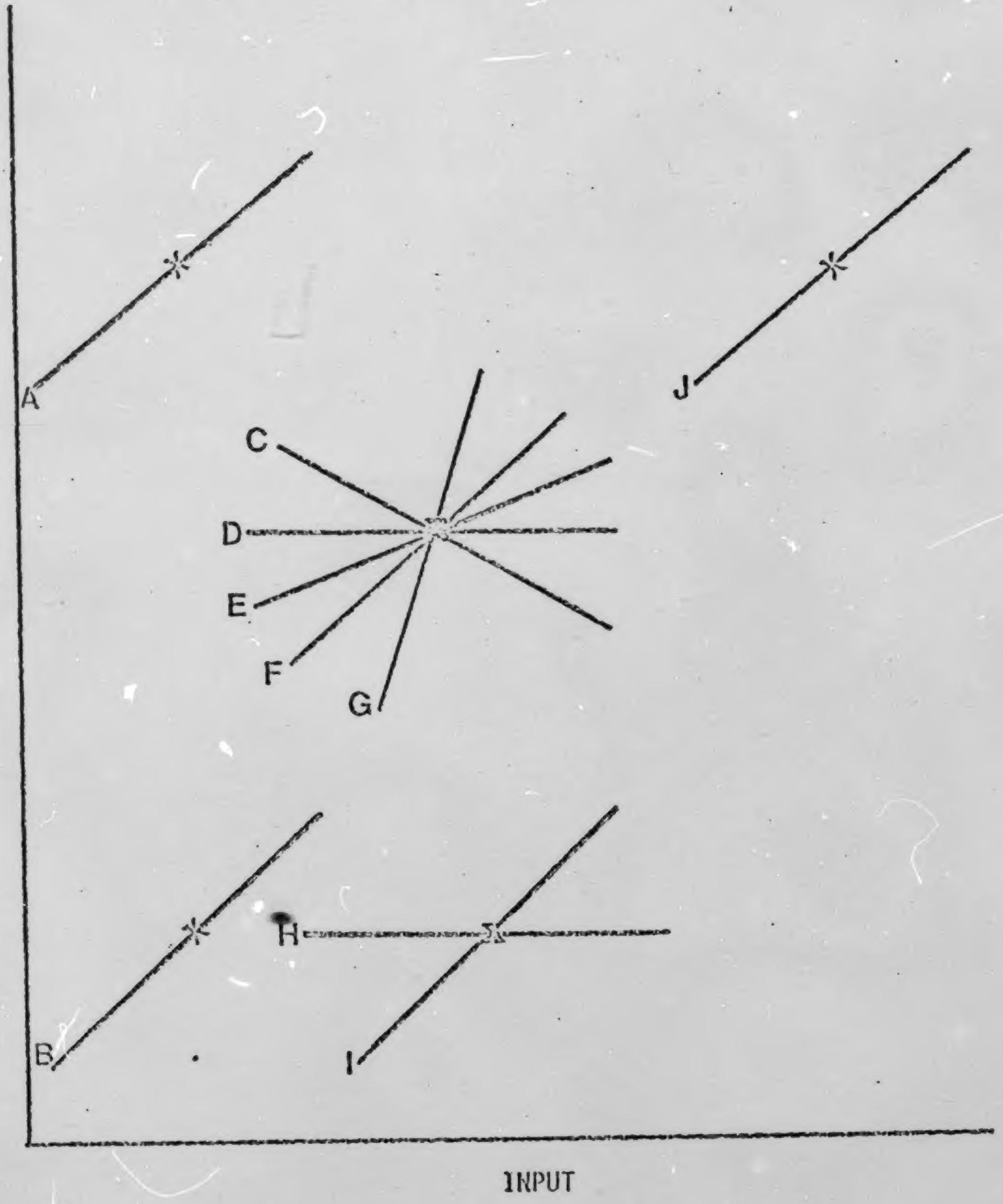

A
J
C
D
E
F
G
H
B
I
INPUT

FIGURE 1. Regression of Outcome (Y) on Input (X) Within Different Classrooms.

indicates the magnitude of the problem they will have to overcome.

The regression of pupil outcome on pupil inputs (which we denoted by β_c) is likely to vary across classrooms. Cronbach (1976, p. 3.6) cites as sources of the variation in β_c: (a) sampling variability due to chance and stability problems due to small class sizes when the processes operating in the classes are the same, (b) differences in the selection factors forming the classes, and (c) differences in causal processes going on in the classrooms. This last source encompasses the possibility that some teachers have meritocratic effects on pupil outcomes while others have compensatory effects or simply that teachers differ in the degree to which they are meritocratic.

If we can rule out chance effects and different selection rules as reasonable explanations, the variation in β_c becomes a potent source of information to researchers and policy makers especially when such information is combined with the adjusted class effects discussed earlier. Ideally, from a social good perspective, the BTES study would hope to identify teaching methods and skills that result in large adjusted mean outcomes with redistributive (compensatory) within-class effects. One can even argue that isolation of teacher attributes and skills that are associated with compensatory or meritocratic within-class effects offers a cost-benefit in terms of the match of pupils, teachers, and schools that cannot be obtained by the consideration of adjusted class effects on mean outcomes.

ALTERNATIVE ANALYSIS MODELS FOR MULTILEVEL DATA

We consider three analysis models for multilevel data -- Cronbach's (1976; Cronbach and Webb, 1975) between-group, pooled within-group analysis, the Keesling-Wiley (1974) two-stage hierarchical analysis, and a modification of the slope-intercept analysis proposed by Burstein (1976b). All three methods respond at least in part to the concerns raised about the choice of units and examine at least two components of pupil outcomes.

Cronbach's Between Group, Pooled Within-Group Analysis.

It is difficult to do adequate justice to Cronbach's detailed documentation of the reasons and methodology for decomposition of overall pupil effects. A succinct statement of his key argument is that the usual overall between-student analysis combines two kinds of relationships -- those operating between collectives (reflected in β_b and adjusted class effects) and those operating among persons within collectives (reflected in β_w and β_c) -- into a composite that is rarely of substantive interest (see, e.g., Cronbach, 1976, pp. 10.3ff.). Cronbach reminds us that β_t, the overall between-student coefficient from the regression of Y_{cp} on X_{cp},

$$Y_{cp} - \bar{Y}_{..} = \beta_t (X_{cp} - \bar{X}_{..}) + U_{cp} ,$$

has been shown by Duncan, Cuzzort, and Duncan (1961, p. 66) to be a composite of β_b and β_w

$$\beta_t = \eta_X^2 \beta_b + (1 - \eta_X^2)\beta_w$$

where η_X^2 is the intraclass correlation or correlation ratio of X. He (1976; Cronbach and Webb, 1975) goes on to recommend that between-group effects and individuals-within-group effects should be examined separately.

In its most parsimonious form, Cronbach would examine the following

Between-Groups: $$\bar{Y}_{c.} - \bar{Y}_{..} = \underset{\sim}{b}_{\bar{Y}_{c.}\underset{\sim}{T}}\underset{\sim}{T} + \beta_b (\bar{X}_{c.} - \bar{X}_{..}) ,$$

where the $\underset{\sim}{b}_{\bar{Y}_{c.}\underset{\sim}{T}}\underset{\sim}{T}$ is the effect of teachers on mean outcomes after controlling for between-class differences in inputs. The $\underset{\sim}{b}_{Y_{c.}\underset{\sim}{T}}\underset{\sim}{T}$ are captured in the adjusted class effects $[\bar{Y}_{c.} - \beta_b (\bar{X}_{c.} - \bar{X}_{..})]$ discussed in the earlier decomposition.

Pooled Within-Groups: $$Y_{cp} - \bar{Y}_{c.} = \beta_w (X_{cp} - \bar{X}_{c.}) .$$

Thus, his primary concerns are with the adjusted collective effects of instruction as reflected by the adjusted class mean outcomes and with the overall redistributive properties of classroom instruction as reflected by the pooled within-class regression, β_w.

The Cronbach report does consider specific within-group regressions (pp. 3.2.ff., pp. 5.5.ff., pp. 8.1.ff. and elsewhere) but mostly indirectly. He provides within-class regression coefficients for classes in both treatment conditions from the Anderson ATI study (pp. 5.5.ff.) and decomposes the variance of pupil outcomes in the fashion described earlier [see p. 5 above, equation (3.5) in Cronbach (1976, p.3.2)]. Cronbach concludes that in Anderson's study, the specific-within-class regressions do not account for much variance (4.1% in the Drill condition and 5.6% in the Meaning condition). Yet, the proportions of outcome variance attributable to adjusted class effects are also small (5.0% for the Drill condition and 6.9% for the Meaning condition). We would argue that both adjusted class effects and specific-within-class effects are potentially meaningful and interpretable or both are not.

Cronbach's examination of aptitude-treatment interaction studies (1976, Chapter 5) is relevant to the argument for examining specific within-class regressions. If each treatment group were represented by a single class, the theory underlying ATI research would consider the comparison of the two pooled within-class regressions as a primary concern. In a sense, specific within-treatment slopes are analogous to specific within-class slopes. To carry the analogy a step further, the classes in the BTES study are treatment groups. Therefore, a search for explanations of differences in within-class slopes in the BTES study can be justified on the same grounds as the consideration of the antecedents of differences in within-treatment slopes in ATI research.

Thus, though Cronbach emphasizes the consideration of between-group and pooled within-group effects, his report clearly recognizes the significance of heterogeneous within-group regressions. At the same time, he offers no strategy for systematic examination of determinants of differences in within-class slopes. In the studies Cronbach considered, differences in within-class slopes that appeared to be significant were generally traceable to anomalous students within classes (e.g., Cronbach, 1976, pp. 5.21-22).

Keesling and Wiley's Hierarchical Analysis

Keesling and Wiley (1974) set out to define a model for disentangling the effects of variables defined solely at the school level from those defined at the level of the pupil. (There is no loss in translating this method to a setting with pupils and classrooms as meaningful levels in the discussion that follows.) The process of disentanglement involved two stages: (1) the adjustment of the effects of individual background characteristics on outcome for the effects of the classrooms in which the individuals receive instruction and (2) adjustment of the effects of classroom characteristics for those of the individuals. The adjusted effect estimates and other byproducts of the adjustment process are relevant to our interest in the components of the effects on pupil outcomes.

To find the adjusted estimates of pupil-level variables, the model must be completely specified at the class level (i.e., all outcome-relevant class variables are measured) and individual effects must be adjusted for all class-level variables in the model. (See Keesling and Wiley, 1974, p. 4.) Keesling and Wiley point out that an adjustment for the interactions of class and student variables should be made if they exist. (Their assumption of the non-existence of such interactions seems to be made more on the basis of complications in the computational algorithm than on theoretical grounds.) The equation for the adjustment of the effects of a single individual-level explanatory variable can be written as:

Adjusted Individual Effects: $$Y_{cp} - \bar{Y}_{c.} = \beta_w(X_{cp} - \bar{X}_{c.}) + \underset{\sim}{\gamma}\underset{\sim}{Z}_c$$

$$+ \underset{\sim}{\beta}_c(\underset{\sim}{Z}_c X_{cp}) + \varepsilon_{cp}$$

where β_w is the pooled within-class slope, the $\underset{\sim}{Z}_c$ are dummy variables for classrooms whose effects are expressed in $\underset{\sim}{\gamma}$, the $\underset{\sim}{Z}_c X_{cp}$ are the interactions of classes with the individual-level variable whose effects are given by the $\underset{\sim}{\beta}_c$. These $\underset{\sim}{\beta}_c$ reflect any heterogeneity of regressions across classes.

If the individual-level variables are adjusted in the fashion described by Keesling and Wiley, estimates of two of the components of interest to us -- the pooled within-class slope and specific within-class slopes -- are provided. β_w is presumably interpretable in the usual manner. Keesling and Wiley do not address the question of how to interpret the information provided by the $\underset{\sim}{\beta}_c$ or how to proceed with the next stage of adjustment when there is significant heterogeneity of regressions. They do state, however, that "non-homogeneity of regression would preclude the use of the same adjustment algorithm for each school" (1974, p. 4).

In the second stage of their analysis, Keesling and Wiley use the adjusted effects of individual-level variables, aggregated over pupils within schools, to determine the adjusted effects of school level variables. In practice they carry out the following:

(1) Determine the pooled within-class slope

$$\hat{Y}_{cp} = \bar{Y}_{c.} + \beta_w(X_{cp} - \bar{X}_{c.}) \quad .$$

(2) Find the mean predicted outcome for each class

$$\hat{\bar{Y}}_{c.} = \beta_w\bar{X}_{c.} \quad .$$

(3) Fit a model at the class level regressing observed class mean outcome on class-level explanatory variables and predicted class mean outcome:

Adjusted Class Effects $\bar{Y}_{c.} = \gamma_o + \underset{\sim}{\gamma}\underset{\sim}{T} + \lambda(\hat{\bar{Y}}_{c.}) + \phi_i \,)$

where $\underset{\sim}{\gamma}$ are the adjusted effects of the class-level variables, denoted here by $\underset{\sim}{T}$, and λ allows partial removal of additional specification bias due to the omission of class-level variables which are correlated with the sum of the average individual-level effect values represented in $\hat{\bar{Y}}_c$. If all relevant class-level variables are included, $\lambda = 1$.

The $\underset{\sim}{\gamma}$ are analogous to the adjusted classroom effects from a between-class analysis in the style of Cronbach. Thus, the Keesling-Wiley analysis provides information relevant to the thrid of the components of interest. And like Cronbach, they exhibit little interest in explanations of heterogeneous within-class regressions.[3]

Analysis of Slopes and Intercepts

In an earlier paper (Burstein, 1976), we expressed the concern that the Keesling-Wiley approach fails to adequately reflect differences among within-class slopes. The paper cited regression slopes such as those depicted in Figure 1 to emphasize the problem. That criticism was overstated, but the fact remains that neither Keesling and Wiley nor Cronbach consciously offer procedures for analyzing the determinants of heterogenity in within-class slopes.

As an alternative to Keesling and Wiley, we suggested the following analysis:

(1) For each class find[4]

Specific Within Class $\bar{Y}_{cp} = \alpha_c + \beta_c(\bar{X}_{cp} - \bar{X}_{c.})$

where α_c and β_c are the adjusted intercept and slope for class c.

(2) Fit a model at the class level with α_c and $\underset{\sim}{\beta}_c$ as outcomes and class-level (or higher level) explanatory variables as independent variables:

$$\alpha_c = \underset{\sim}{b}_{\alpha_c T}\underset{\sim}{T} + \tau_1 \quad ,$$

$$\beta_c = \underset{\sim}{b}_{\beta_c T}\underset{\sim}{T} + \tau_2 \quad .$$

In the above, the adjusted class or teacher effects are expressed in $\underset{\sim}{b}_{\beta_c T}$. The regression of the slopes on teacher effects presumably tells us about the antecedents of any heterogeneity of within-class regressions. For example, the values $\underset{\sim}{b}_{\beta_c T}$ reflect systematic effects of class characteristics on slopes. This interpretation of $\underset{\sim}{b}_{\beta_c T}$ is directly relevant to an elaboration of the components of the effects on pupil outcomes.

Whereas Cronbach and Keesling and Wiley deemphasize specific within-class slope analyses, the slope-intercept approach emphasizes the specific within-class effects (class means and slopes) to the partial exclusion of consideration of the pooled within-class regression. The analysis of pooled within-slope represents the fall-back position in case the analysis of the $\underset{\sim}{\beta}_c$ produces no interpretable antecedents of slope heterogeneity.

It is important to reiterate that the distinctions among the alternative methods for analyzing multilevel data described above are a matter of emphasis. Adjusted class effects, the pooled within-group regressions and specific within-group regressions are either estimated directly or can be determined within the framework of all three approaches.

We turn next to preliminary results from a computer-simulated educational effects model where random variation in β_c or systematic variation related to teacher effects have been introduced. We compare only the Keesling-Wiley and the slope-intercept methods. The between-group analysis recommended by Cronbach is analogous to the Keesling-Wiley estimation of the adjusted teacher effects, $\underset{\sim}{\gamma}_T$, and to the slope-intercept estimation of $\beta_{\alpha_c T}$. His pooled within-group regression is a byproduct of the Keesling-Wiley estimates of adjusted class effects and is of minimal importance to the slope-interpreted method though it can be generated as part of the analysis.

COMPUTER SIMULATION OF EDUCATIONAL EFFECTS

Linn developed a computer program to simulate a structural model of educational effects involving an outcome, a pretest, a school quality index, a teacher quality index, and family background. The model can be represented schematically as

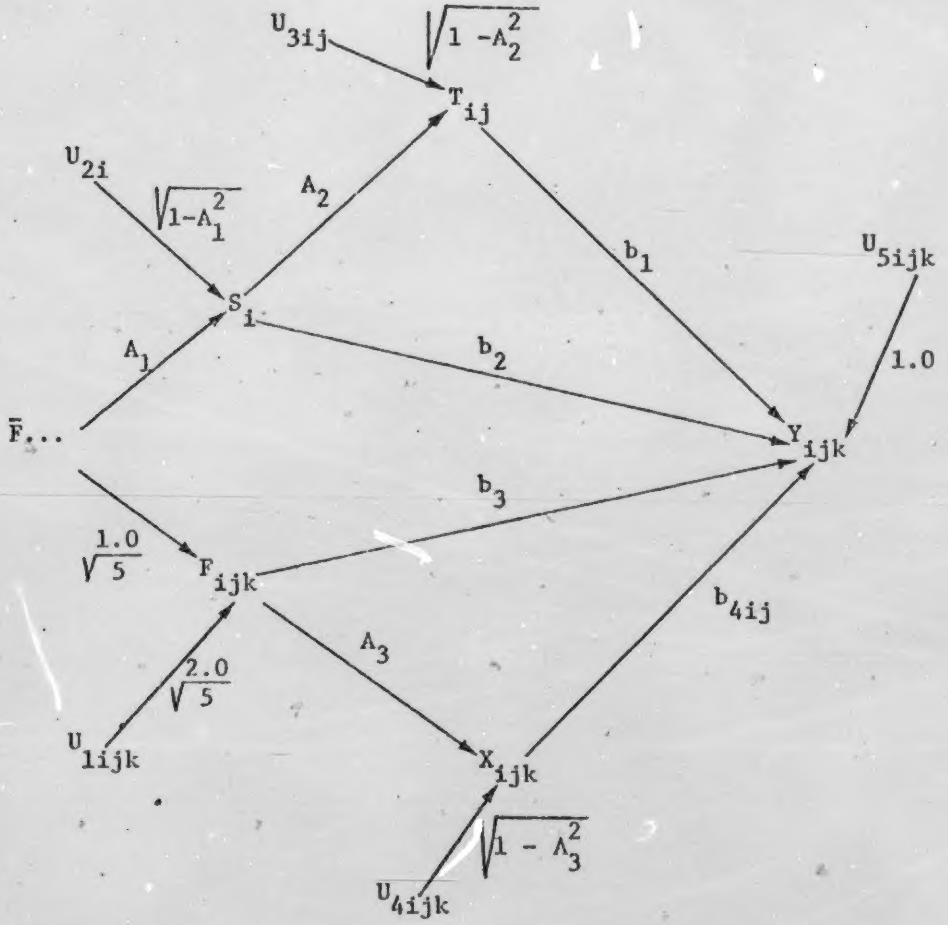

U_{3ij}
$\sqrt{1-A_2^2}$
T_{ij}
U_{2i}
$\sqrt{1-A_1^2}$
A_2
b_1
U_{5ijk}
S_i
A_1
b_2
1.0
$\bar{F}...$
Y_{ijk}
b_3
$\frac{1.0}{\sqrt{5}}$
F_{ijk}
b_{4ij}
A_3
$\frac{2.0}{\sqrt{5}}$
U_{1ijk}
X_{ijk}
$\sqrt{1-A_3^2}$
U_{4ijk}

Figure 2

Simulation Model

The U's in Figure 1 all represent disturbance terms in the model. They are all obtained as pseudo random numbers from independently distributed normal distribution with means and variance chosen for convenience.

Thus,

$$U_{1ijk} \sim N(\sqrt{5,1}\,),$$
$$U_{2i} \sim N(5,1),$$
$$U_{3ij} \sim N(5,1),$$
$$U_{4ijk} \sim N(5,1),$$

and

$$U_{5ijk} \sim N(0,1).$$

The remaining symbols in Figure 1 represent characteristics of substantive interest which are to be simulated or parameters of the model.

Specifically, the model states that:

The posttest performance for student k, in class j and school i, Y_{ijk}, is a function of the student's pretest score, X_{ijk}, the student's family background, F_{ijk}, the school quality, S_i, the teacher quality, T_{ij}, and a random disturbance term, $U_{5ijk}(\sim N(0,1))$. The structural form of the model is:

$$Y_{ijk} = b_1 S_1 + b_2 T_{ij} + b_3 F_{ijk} + b_{4ij} X_{ijk} + U_{5ijk} \,. \qquad (1)$$

The b_1, b_2, b_3 and b_{4ij} are model parameters which may be selected to reflect various degrees of dependence of posttest on the other variables in the model. It should be noted that b_{4ij} may vary across classrooms. Furthermore, as outlined below, the b_{4ij} may be a random or systematic function of teacher quality.

Values of the variables in the model, which eventually determine posttest, i.e,, Y_{ijk}, are obtained as follows:

(1) Mean family background for school i is obtained by generating a pseudo random number from a normal distribution with a mean of 5 and a standard deviation of 1 .

(2) School quality (S_i) is, in part, determined by the average family background of its students:

$$S_i = A_1\bar{F}_{i..} + \sqrt{1 + A_1^2}\,U_{2i} , \qquad (2)$$

where A_1 is a parameter of the model (to be chosen between -1 and +1) which controls the degree of dependence of school quality on family background.

(3) Teacher quality (T_{ij}) is a function of school quality and a random disturbance,

$$T_{ij} = A_2S_i + \sqrt{1 - A_2^2}\,U_{3ij} , \qquad (3)$$

where A_2 is a parameter of the model ($-1 \leq A_2 \leq 1$).

(4) Family background for a student is obtained from the mean family background of the student's school ($\bar{F}_{i..}$) plus a disturbance, i.e.,

$$F_{ijk} = (1/\sqrt{5})\bar{F}_{i..} + (2/\sqrt{5})U_{1ijk} . \qquad (4)$$

The coefficients in (4) were chosen to yield a convenient variance and to control the relative size of the between- and within-school variances on the family background variable.

(5) Finally, a student's pretest is obtained as a function of that student's family background, and a disturbance term,

$$X_{ijk} = A_3F_{ijk} + \sqrt{1 - A_3^2}\,U_{4ijk} , \qquad (5)$$

where A_3 is a model parameter which determines the degree of relationship between F and X.

As noted above, the value of b_{4ij} varies from classroom to classroom. The variation in the b_{4ij} may be either random or systematically related to teacher quality. More explicitly,

$$b_{4ij} = b_4 + c(.1)U_{3ij} + \sqrt{1 - c^2}U_{6ij} , \qquad (6)$$

where U_{3ij} is the pseudo random number used to determine teacher quality (see equation 3), U_{6ij} is a pseudo random number from a normal distribution with mean zero and standard deviation of 0.1, and c is a model parameter which determines the relationship between teacher quality and slope. The mean of the b_{4ij} is b_4 and the standard deviation is 0.1.

The analyses and discussion that follow are based on the generation of a single set of pseudo normal deviates to simulate a sample of $n_i = 30$ schools with $n_j = 3$ classrooms per school and $n_k = 30$ pupils per classroom $(N = n_i n_j n_k = (2700)$. Twelve distinct sets of data were formed by varying school quality ($b_1 = .0, .2$), teacher quality ($b_2 = .0$ and $.3$ when $b_1 = .0$; $b_2 = .3$ when $b_1 = .2$) and the effects of teacher quality on the variation in b_{4ijk}($c = .4, 0, -.4, -.8$) while fixing the remaining parameters ($b_3 = .2$, $b_4 = 1.0$, $A_1 = .2$, $A_2 = .1$ and $A_3 = .4$).

In its simplist form (no school and teacher effects and random variation in b_{4ij}), (1) becomes

$$Y_{ijk} = .2F_{ijk} + 1.0X_{ijk} + U_{5ijk} \quad .$$

This version of model (1) reflects an extremely pessimistic view of the effects of schooling. Any apparent effects of school and teacher quality are spurious and can be explained by the tendency of school quality to be positively associated with mean family background and for better schools to have better teachers. Moreover, any teacher-to-teacher deviations from the average relationship of pretest to outcome are merely random fluctuations. No teacher is systematically better or worse at redistributing the rank orderings of measured skills within his or her class.

The opposite extreme, which reflects an egalitarian's idea of the effective educational process, is represented by

$$Y_{ijk} = .2S_i + .3T_{ij} + .2F_{ijk} + [1 - (.8)(.1)U_{3ij}]X_{ijk} + U_{5ijk} \quad .$$

That is, $b_1 = .2$, $b_2 = .3$ and c in equation (6) is set at -.8. In this case, both school quality and teacher quality affect pupil outcomes over and above the effects of background and ability. In addition, the better teachers are compensatory; that is, they produce a larger boost in performance for their lower ability students.

We have also generated data in which better teachers produce a larger boost in performance for their more able students, thereby reproducing the inequities in performance that pupils bring to the class. In this version of model (1), the meritocratic model, the effect of teacher quality on variation in the within-class slope is positive ($c = .4$).

Table 1 provides the means and between-student and between-class standard deviations for the regressors appearing in the simulation. The computer simulated means and standard deviations at the pupil level deviate from their expected values, which for the means are 5, 5, 2.24, and 4.58 for S, T, F, and X respectively. The expected values of the between-student variances are 1.0 for all four variables. The discrepancies between the observed and expected values of these means and variances reflect the limitation of examining a single sample of pseudo values.

Table 1

Table 2 contains the correlations among the regressors in (1) at both pupil and class levels. The values of these correlations are fixed over all combinations of models generated from (1). Several of the between-student correlations are lower than might be expected. (E.g., r_{FS} might be thought to be around .3; r_{SX} might be expected to be greater than .2). At the class level, the value of $r_{\bar{F}\bar{X}}$ is usually large given the simulated value of r_{FX}. Despite these problems, the overall pattern of correlations at both student and class levels is reasonable for our purposes.

Table 2

Table 1. Means and between-student and between-class standard deviations of school quality (S), teacher quality (T), pupil's family background (F), and pretest (X) for the simulated educational effects model.

Variable	Mean	Between-Student S.D. (N=2700)	Between-Class S.D. (N=90)
S	5.060	.872	.877
T	5.576	.871	.876
F	2.091	1.032	.494
X	4.710	.891	.257

Table 2. Between-student and between class correlations of school quality (S), teacher quality (T), family background (F), and pretest (X).

		S	T	F	X	
	S		.11	.32	.23	**UPPER TRIANGLE:** Between-Class Correlations
	T	.12		.05	.03	
LOWER TRIANGLE: Between-Student Correlations	F	.16	.03		.78	
	X	.08	.02	.48		

The means and between-student and between-class standard deviations of outcome (Y) are provided in Table 3 for all combinations of simulated conditions. The effects of the successive introduction of direct teacher and then direct school effects on the means and standard deviations at the pupil level are evident and expected.

Table 3

Though the variation in teacher effects on within-class slopes has a minor impact on the pupil-level parameters, it apparently influences between-class variability in outcome in a significant and complex fashion. In the case of random variation in within-class slopes, the between-class variability of Y increases as teacher effects ($b_{YT} = .3$) and school effects ($b_{YS} = .2$) are successively added. The same pattern appears with meritocratic effects on slopes ($c = .4$) but the magnitudes of differences in standard deviations between the teacher and school effects conditions (i.e., $b_{YS} = .0$, $b_{YT} = .0$ v.s. $b_{YS} = .0$, $b_{YT} = .3$ vs. $b_{YS} = .2$, $b_{YT} = .3$) change from the random slope variation case.

The behavior of the compensatory models ($c = -.4$ and $c = -.8$) suggests that teacher effects on slopes and on mean outcomes behave as opposing forces. The inclusion of both types of effects reduces between-class variability below what the variability would be if either one or the other were present. The inclusion of school effects on outcome in addition to the teacher effect ($b_{YS} = .2$, $b_{YT} = .3$) is sufficient to cancel out the smaller compensatory effect ($c = -.4$) and raise the variability beyond the case where neither teacher nor school effects are present ($b_{YS} = .0$, $b_{YT} = .0$) but is not strong enough to overcome the larger compensatory effect ($c = -.8$).

Table 3. Mean Outcome ($\bar{Y}...$) and between-student and between-class standard deviation of outcome (Y) as a function of the effects of S and T on Y and of T on the within-class regression of Y on X.

Effect of Teacher Quality on variation in β_c	b_{YS}	b_{YT}	$\bar{Y}...$	$\hat{\sigma}_{Y_{BS}}$ *	$\hat{\sigma}_{Y_{BC}}$ **
.4	.0	.0	5.194	1.492	.600
	.0	.3	6.867	1.538	.710
	.2	.3	7.878	1.565	.765
.0	.0	.0	5.177	1.499	.622
	.0	.3	6.850	1.513	.658
	.2	.3	7.862	1.542	.723
-.4	.0	.0	5.156	1.503	.637
	.0	.3	6.828	1.486	.599
	.2	.3	7.840	1.518	.673
-.8	.0	.0	5.128	1.500	.637
	.0	.3	6.801	1.454	.521
	.2	.3	7.813	1.485	.603

* $\hat{\sigma}_{Y_{BS}}$ is the between-student standard deviaiton of Y.

** $\hat{\sigma}_{Y_{BC}}$ is the between-class standard deviation of Y.

The between-student and between-class correlations of regressors with outcomes are presented in Table 4. At the student level, the correlations of background (F) and pretest (X) with outcome (Y) are relatively stable across the conditions of the simulation. The correlations of school quality (S) with outcome (Y) exhibit changes as teacher and school effects are successively added that are consistent across the four slope variability conditions ($c = .4, 0, -.4, -.8$). The between-class correlations of S, F, and $\bar{X}$ with $\bar{Y}$ behave much like the between-student correlations of S, F, and X with Y.

Table 4

The most apparent anomalies in Table 4 are the negative relationships between teacher quality and outcome at the class level when teachers have mildly compensatory effects on within-class slope, β_c ($c = -.4$), and at both class and student levels when the compensatory effects are stronger ($c = -.8$). Taking the negative correlations at face value, one might reach the unrealistic conclusion that better teachers yield classes with lower mean outcomes in cases where the effect of teacher quality on pupil outcome is specified to be positive (E.g., when $b_{YS} = .2$, $b_{YT} = .3$ and the teacher effect on variation in β_c is $-.8$). This suggests that a between-class analysis can lead us to identify the wrong teachers as "effective" if good teachers are compensatory.

The pupil-level and class-level regressions of outcome on pretest, family background, school quality and teacher quality (Tables 5 and 6) provide further support for our concern about systematic class-to-class variations in slopes. When variation in the within-class regression of outcome on pretest is random, the parameter estimates at the pupil level deviate from the specified value only slightly.[5] At the class level $b_{\bar{Y}S}$, $b_{\bar{Y}T}$, and $b_{\bar{Y}\bar{F}}$ are also accurately estimated for random perturbations in the within-class slope.

Tables 5 - 6

Table 4. Between-student and between-class correlations of Y with S, T, F and X for varying effects of S and T on Y and T on β_c.

Effect of T on Variation in β_c	$b_{YS}=$	$b_{YT}=$	CORRELATIONS WITH Y							
			BETWEEN STUDENT				BETWEEN CLASS			
			S	T	F	X	S	T	F	X
.4	.0	.0	.09	.10	.40	.65	.20	.24	.43	.46
	.0	.3	.10	.26	.40	.63	.21	.57	.38	.40
	.2	.3	.21	.27	.41	.63	.42	.55	.43	.43
.0	.0	.0	.10	-.02	.40	.64	.22	-.07	.41	.44
	.0	.3	.11	.15	.40	.64	.26	.34	.41	.43
	.2	.3	.22	.16	.41	.63	.48	.33	.45	.45
-.4	.0	.0	.10	-.14	.40	.64	.23	-.34	.41	.45
	.0	.3	.12	.03	.41	.65	.30	.07	.46	.49
	.2	.3	.23	.04	.42	.64	.53	.09	.49	.49
-.8	.0	.0	.10	-.25	.40	.64	.23	-.61	.44	.48
	.0	.3	.12	-.08	.42	.66	.34	-.24	.56	.60
	.2	.3	.24	-.07	.43	.66	.58	-.17	.58	.59

Table 5. Between-student regressions of outcome (Y) on school quality (S), teacher quality (T) family background (F) and pretest (X) as a function of the relation of school and teacher quality to outcome and of teacher quality to variations in within-class slopes (β_c).

Effect of Teacher Quality on Variation in β_c	Direct Effects of S and Y on					
	b_{YS}	b_{YT}	$\hat{b}_{YS}$	$\hat{b}_{YT}$	$\hat{b}_{YF}$ [a]	$\hat{b}_{YX}$ [a]
.0	.0	.0	.07*	-.07*	.17*	.98*
	.0	.3	.07*	.23*	.17*	.98*
	.2	.3	.27*	.23*	.17*	.98*
-.8	.0	.0	.12*	-.47*	.18*	.84*
	.0	.3	.12*	.17*	.18*	.84*
	.2	.3	.32*	.17*	.18*	.84*

*Parameter estimates which exceed twice their standard errors.

[a] The parameters in the simulation are $b_{YF} = b_3 = .2$ and $b_{YX} = b_4 = 1.0$.

Table 6. Between-class regressions of outcome ($\bar{Y}$) on school quality (S), teacher quality (T), family background ($\bar{F}$) and pretest ($\bar{X}$) as a function of the relation of school and teacher quality to outcome and teacher quality to variation of within-class slopes (β_c).

Effect of Teacher Quality on Variation in β_c	Direct Effects of S and Y on					
	b_{YS}	b_{YT}	$\hat{b}_{\bar{Y}S}$	$\hat{b}_{\bar{Y}T}$	$\hat{b}_{\bar{Y}\bar{F}}$[a]	$\hat{b}_{\bar{Y}\bar{X}}$[a]
.4	.0	.0	.04	.15*	.17	.78*
	.0	.3	.04	.45*	.17	.78*
	.2	.3	.23*	.45*	.17	.79*
.0	.0	.0	.08	-.07	.17	.75*
	.0	.3	.08	.23*	.17	.75*
	.2	.3	.28*	.23*	.17	.75*
-.4	.0	.0	.12*	-.27*	.18	.77*
	.0	.3	.12*	.03	.18	.77*
	.2	.3	.32*	.03	.18	.77*
-.8	.0	.0	.12*	-.47*	.20	.84*
	.0	.3	.12*	-.17*	.20	.84*
	.2	.3	.32*	-.17*	.20	.84*

*Parameter estimates which exceed twice this standard errors.

[a] The parameters of the simulation for background and pretest are $b_{YF} = b_3 = .2$ and $b_{YX} = b_4 = 1.0$.

When teachers have either compensatory or meritocratic effects on within-class slopes, the picture becomes more complicated. Though the estimates of b_{YF} and b_{YX} are virtually unaffected, the estimates of b_{YS} are mildly inflated and the estimates of b_{YT} are dramatically distorted in the direction of the effects of teacher quality on the slope.

Our main conclusion from the simulation at this point is that neither pupil-level nor class-level regression analyses yield the correct estimates of school and teacher effects when there are systematic differences in within-class regressions of outcome on pretest. Irrespective of the chosen level of analysis, the investigator must systematically examine the within-class regressions in order to detect that his effect estimates are erroneous.

RESULTS FROM SLOPE-INTERCEPT AND KEESLING-WILEY ANALYSES

Estimates of within-class slopes (β_c) and intercepts (α_c) were determined for each class for the twelve combinations of effects in the simulation. In the analyses that follow, the intercepts are adjusted by adding $\hat{\beta}_c\bar{X}...$ to modify the strong negative correlation between $\hat{\alpha}_c$ and $\hat{\beta}_c$. So

$$\hat{\alpha}_c = \hat{\alpha}_{c(u)} + \hat{\beta}_c\bar{X}...$$

where $\hat{\alpha}_{c(u)}$ is the unadjusted intercept and

$$E(\hat{\alpha}_c) = \bar{Y}... \quad .$$

Table 7 contains the means and standard deviations for the adjusted $\hat{\alpha}_c$ and the $\hat{\beta}_c$ and their intercorrelations over the 90 classrooms from each version of the model. There are inconsequential deviations, attributable to rounding error, of the means of $\hat{\alpha}_c$ from the $\bar{Y}...$ reported in Table 3. The mean of $\hat{\beta}_c$ is constant for a given effect of teacher quality on variation in within-class slopes and approximately equal over all conditions of the simulation. The between-class standard deviation of $\hat{\beta}_c$ is essentially constant for all conditions and differs from the .1 value specified for b_{4ij} in the simulation because the latter is the standard deviation of a partial regression coefficient.

Table 7

Table 7. Class means and standard deviations and correlations of α_c^a and β_c from twelve combinations of teacher (T) and school (S) effects on pupil outcomes and teacher effects on within-class slopes. (N = 90 classrooms)

Effect of Teacher Quality on Variation	Direct Effect of S and T on Y		$\hat{\alpha}_c$		$\hat{\beta}_c$		
in β_c(b)	b_{YS}	b_{YT}	Mean	S.D.	Mean	S.D.	$r_{\hat{\alpha}_c \hat{\beta}_c}$
.4	.0	.0	5.198	.537	1.082	.234	.301
	.0	.3	6.870	.654	1.082	.234	.313
	.2	.3	7.882	.700	1.082	.234	.286
.0	.0	.0	5.181	.564	1.078	.234	.308
	.0	.3	6.854	.601	1.078	.234	.287
	.2	.3	7.866	.655	1.078	.234	.262
-.4	.0	.0	5.159	.577	1.074	.234	.314
	.0	.3	6.832	.533	1.074	.234	.260
	.2	.3	7.844	.597	1.074	.234	.233
-.8	.0	.0	5.130	.566	1.068	.235	.314
	.0	.3	6.803	.427	1.068	.235	.223
	.2	.3	7.815	.505	1.068	.235	.190

a. The $\hat{\alpha}_c$ are the adjusted intercepts found by adding $\hat{\beta}_c\bar{X}$... to the computer-generated values. The adjustment was made to reduce the strong negative correlation between $\hat{\alpha}_c$ and $\hat{\beta}_c$.

b. The values .4, .0, -.4, and -.8 are the direct effects of T on variation in β_c. Since the standard deviation of β_c in the simulation model was set at .1, the magnitudes of the effects can be expected tobe .04, .00, -.04 and -.08.

c. The standard deviation reported for $\hat{\beta}_c$, the within-class regressions, differ from the value of .1 set as the standard deviation of the partial regression coefficient b_4 in the simulation model.

The slope-intercept approach uses the $\hat{\alpha}_c$ and $\hat{\beta}_c$ from the simulated data as dependent variables with teacher and school quality as independent variables in an analysis at the class level. The estimates of interest are $\hat{b}_{\hat{\alpha}_c T}$, $\hat{b}_{\hat{\alpha}_c S}$, $\hat{b}_{\hat{\beta}_c T}$ and $\hat{b}_{\hat{\beta}_c S}$ from the equations

$$\hat{\alpha}_c = \mu_1 + b_{\hat{\alpha}_c T}T + b_{\hat{\alpha}_c S}S + \epsilon_1 \qquad (7a)$$

$$\hat{\beta}_c = \mu_2 + b_{\hat{\beta}_c T}T + b_{\hat{\beta}_c S}S + \epsilon_2 \qquad (7b)$$

In order to illustrate the Keesling-Wiley technique, the class mean predicted outcome, $\hat{\bar{Y}}_c$, is estimated using the unadjusted $\hat{\alpha}_c$ and the $\hat{\beta}_c$:

$$\hat{\bar{Y}}_c = \hat{\alpha}_{c(u)} + \hat{\beta}_c \bar{X}_{ij} . \qquad (8)$$

Next, the observed class mean outcomes, $\bar{Y}_c$, are regressed on T and $\hat{\bar{Y}}_c$:

$$\bar{Y}_c = \mu_3 + \gamma_T T + \gamma_{\hat{\bar{Y}}_c} \hat{\bar{Y}}_c + \epsilon_3 . \qquad (9)$$

In (9), γ_T represents the direct effects of T on $\bar{Y}_c$ and any deviation of $\gamma_{\hat{\bar{Y}}}$ from 1 reflects specification errors in the class-level model (Keesling and Wiley, 1974, p. 8).[6]

Table 8 contains the necessary information for comparing estimates of teacher effects from the slope-intercept analysis with the estimates from the Keesling-Wiley approach. Table 9 provides the estimates of the school quality effects from the slope-intercept analyses.

Tables 8 - 9

Since the adjusted α_c are closely related to class means on Y, the effects of T on the α_c are quite similar to the effects of T, γ_T, as estimated from the Keesling-Wiley analysis. In fact, by examining Tables 5 and 6 as well as Table 8, we see that standard between-student class analyses give the same estimates of teacher effects on outcomes as the slope intercept analysis and are similar to the Keesling-Wiley approaches, i.e.,

Table 8. Comparison of estimates of effects of teacher quality (T) on within-class slopes (β_c) and intercepts (α_c)[a] with a Keesling-Wiley analysis of effects of teacher quality (T) and predicted outcomes ($\hat{\bar{Y}}_c$) on observed class means ($\bar{Y}_c$) for different models of teacher (T) and school (S) effects on pupil outcomes and teacher effects on within-class slopes.

Effect of Teacher Quality on Variation in β_c[b]	Direct Effect of School and Teacher Quality on Y		Parameter Estimates for Effects of Teacher Quality on Intercepts and Slope		Keesling-Wiley Estimates	
	b_{YS}	b_{YT}	$b_{\hat{\alpha}_c,T}$	$b_{\hat{\beta}_c,T}$	$\hat{Y}_T$	$\hat{Y}_{\hat{\bar{Y}}}$
.4	.0	.0	.15	.05	.15	.99
	.0	.3	.45*	.05	.45	.99
	.2	.3	.45*	.05	.47	1.13
.0	.0	.0	-.07	.00	-.06	1.00
	.0	.3	.23*	.00	.24	1.00
	.2	.3	.23*	.00	.26	1.14
-.4	.0	.0	-.27*	-.04	-.26	1.05
	.0	.3	.03	-.04	.04	1.05
	.2	.3	.03	-.04	.06	1.19
-.8	.0	.0	-.47*	-.09*	-.45	1.15
	.0	.3	-.17*	-.09*	-.15	1.15
	.2	.3	-.17*	-.09*	-.13	1.30

a. The within-class intercepts, α_c, are adjusted be adding $\beta_c\bar{X}$... to the computer-generated values to remove the strong negative relationship between α_c and β_c in the simulation. In cases where both adjusted and unadjusted α_c were used as dependent variables, the parameter estimates for the effects of teacher quality on α_c differed uniformily by .01 (e.g., when S=0, T=0 and variation in β_c is random, $b_{\alpha_c T}$ = -.06 when the α_c were unadjusted).

b. The values .4, .0, -.4, and -.8 are the direct effects of T on variation in β_c. Since the standard deviation of β_c in the simulation model was set at .1, the magnitudes of the effects can be expected to be .04, .00, -.04 and -.08.

* Parameter estimates which exceed twice their standard errors.

Table 9. Estimates of effects of school quality (S) on within-class slopes (β_c) and intercetps (α_c)[a] for different models of teacher (T) and school (S) effects on pupil outcomes and teacher effects on within-class slopes.

Effect of Teacher Quality on Variation in β_c	Direct Effect of S and T on Y		$b_{\hat{\alpha}_c S}$	$b_{\hat{\beta}_c S}$
	b_{YS}	b_{YT}		
	.0	.0	.05	-.01
.04	.0	.3	.05	-.01
	.2	.3	.25*	-.01
	.0	.0	.10	.00
.00	.0	.3	.10	.00
	.2	.3	.30*	.00
	.0	.0	.13*	.01
-.04	.0	.3	.13*	.01
	.2	.3	.33*	.01
	.0	.0	.15*	.01
-.08	.0	.3	.15*	.01
	.2	.3	.35*	.01

[a] The $\hat{\alpha}_c$ are the adjusted intercepts found by adding $\hat{\beta}_c \bar{X}$... to the computer-generated values. The adjustment was made to reduce the strong negative correlation between $\hat{\alpha}_c$ and $\hat{\beta}_c$.

*Parameter estimates whichexceed twice their standard errors.

$$\hat{b}_{YT} = \hat{b}_{\bar{Y}T} = b_{\hat{\alpha}_c T} \approx \hat{\gamma}_T \quad . \tag{10}$$

Thus, all four methods of analyses over-estimate the direct effects of teacher quality on outcomes in the meritocratic models ($c = .4$) and underestimate the effects of teacher quality in the compensatory models ($c = -.4$ or $-.8$).

The potential advantage of either the slope-intercept analysis or the Keesling-Wiley analysis when compared to the between-student and the between-class regressions is derived from the additional information provided about the adequacy of the analysis of class means. A comparison of the values $\hat{\gamma}_{\hat{\bar{Y}}_c}$ when $b_{YS} = .2$ with its values when $b_{YS} = .0$ indicates that the Keesling-Wiley analysis is highly sensative to the exclusion of the direct effects of school quality from the class-level model. Furthermore, the deviations of $\hat{\gamma}_{\hat{\bar{Y}}_c}$ from 1.0 when $b_{YS} = .0$, suggest, perhaps, that the Keesling-Wiley approach is sensitive to the problem of systematic variation in within-class slopes though apparently such effects must be of substantial magnitude to be detectable. Even so, in the presence of specification errors such as the exclusion of a relevant variable and the existence of large systematic differences in within-class slopes, the Keesling-Wiley techniques sufficiently warn the investigator that caution is necessary in making inferences from this model. The search for additional causes and the reconsideration of adequacy of the pooled within-class adjustment are natural next steps for the forewarned investigator.

In contrast to the Keesling-Wiley analysis and the regressions at the pupil and class levels, the investigator can learn about the direction in which estimates of teacher effects are biased by examining $\hat{b}_{\hat{\beta}_c T}$ from a slope-intercept analysis. For every combination of systematic slope differences in the simulation, $b_{\hat{\alpha}_c T}$ deviates from $b_{YT}(=b_2)$ in the direction of the effect of teacher quality on within-class slopes. When $c = .4$, $b_{\hat{\alpha}_c T}$ overestimates

b_{YT} and when c = 0.4 or -.8, $b_{\hat{\alpha}_c T}$ underestimates b_{YT}. The values .4, .0, -.4, and -.8 are the direct effects of T on variation in β_c. Since the standard deviation of β_c in the simulation model was set at .1, the magnitudes of the effects can be expected to be .04, .00, -.4 and -.08. Thus, as can be seen, $\hat{b}_{\hat{\beta}_c T}$ successfully captures the magnitude and direction of the effects of T on variation in β_c even when the effects are small. That is, estimates of .05, .00, -.04 and -.09 are obtained for the four expected values of .04, .00, -.04 and -.08, respectively.

The results from the slope-intercept analysis are not uniformily positive, however. Unfortunately, we chose two values for c (= .4 and = -.4) in which the effects of T on β_c might be considered nonsignificant by a standard hypothesis test. It is also troublesome that we cannot offer an explaration as to why the systematic effects we induced caused biases in $b_{\hat{\alpha}_c T}$ on the order of 4.5 times the effect of T on β_c.

After adjusting for sampling variability in the specific normal deviates selected, the results from the slope-intercept approach for detecting the effects of school quality exhibited in Table 9 appear to be uniformly satisfactory. In each case, the bias due to sampling variability is at least twice as large as the bias attributable to systematic teacher effects. The small values of $b_{\hat{\beta}_c S}$ would cause the investigator to disregard the effects of systematic differences in within-class slopes as a source of bias in the estimates of the effects of school quality.

A word of caution about the above interpretation needs to be made. The apparent success in detection of school quality effects may be highly illusory. The systematic differences in teacher effects on within-class slopes is a "class-level" effect. By our choice of parameters, we have distributed good and poor teachers in an almost random fashion across schools (r_{ST} = .11 at the between-class level). Therefore, we might expect to find that the average within-class slope from a school is relatively unaffected by school quality. The results from an alternative form for the simulation might have been very different.

CONCLUSIONS AND CAVEATS

Though multiple sources of complications in the analysis of multilevel educational data were described, this paper has focussed on a single concern -- the determination of a method of analysis that will provide accurate estimates of educational effects when there are systematic differences in within-class regressions of outcome on input. Approaches suggested by Cronbach (1976; Cronbach and Webb, 1975) and by Keesling and Wiley (1974) were considered in addition to an analysis utilizing within-class slopes and intercepts as outcomes as recommended in Burstein (1976).

A model of educational effects was simulated wherein the effects of teacher quality on the variation in the within-class slopes was systematically varied over a range of meritocratic (positive relation of teacher quality to within-class slope), random and compensatory (negative relation of teacher quality to within-class slope) effects. The direct effects of teacher quality and school quality on pupil outcome were also varied. Between-student and between-class regressions of outcome on all variables specified in the model were run in addition to the Keesling-Wiley and slope-intercept analyses. For the conditions reflected in the simulation, the following conclusions were reached regarding the identification of educational effects in the presence of systematic differences in between-class slopes:

(1) An overall between-student analysis, a between-class analysis, a Keesling-Wiley analysis and slope-intercept analysis all yield the same misleading estimates of the magnitude of teacher effects on mean class outcome.

(2) The Keesling-Wiley method of analysis provides some indication of misspecification due to deletion of relevant causes and perhaps due to systematic differences in within-class slopes.

(3) The slope-intercept method of analysis provides an indication of the direction of the bias in estimating teacher effects on mean class outcomes and may also suggest the severity of such biases.

The generality of our conclusions is severely limited by the restrictive nature of the simulation we carried out. We considered a relatively straightforward educational effects model generated from a single set of normal deviates whose only significant educative feature were the introduction of systematic teacher effects on within-class slopes. Moreover, the systematic effects we introduced were of the simplist form possible -- direct linear relations of teacher quality to variation in the within-class slopes. It is highly unlikely that actual differences in within-class slopes would arise in this fashion. Any judgments about the likely occurrence of biases as large or larger than those from the simulation when differences among within-class slopes are attributable to more complex sets of systematic factors would be highly suspect.

The findings cited above are not very profound even within the confines of the conditions simulated. It is arguable that a standard test of parallism of regression slopes as in the analysis of covariance will provide the same information generated by the Keesling-Wiley and slope-intercept analyses. Once heterogeneity of the within-class regressions is indicated, an examination of plots of within-class slopes versus variables like teacher quality (which may be unmeasurable) can lead to partial disentanglement of the teacher effects on means from those on slopes. The investigator might also consider a model at the pupil level which examines the effects of teacher quality controlling for differences in within-class slopes. Hannan, Rogosa and Young (personal communication) have suggested this strategy as a viable alternative for overcoming the heterogeneity of regression problems.

In an earlier paper (Burstein, 1976) we argued that the development of approaches for the analysis of multilevel data is essential if we are to avoid looking at effects one class at a time. Cronbach (1976) has stated this same concern more eloquently and with more caution about the possibility of developing a universally successful strategy. If the results from our simulation of a fairly restrictive set of conapplications provides any indication of the difficulties that will be encountered in realistic settings, Cronbach may yet prove to have been overly optimistic. We may eventually have to conclude that the identification of educational effects can best be accomplished by accumulating evidence from case studies of one classroom at a time.

FOOTNOTES

[1] The research reported here was conducted under National Institute of Education Contract NIE-C-74-0123 to Vasquez Associates, Ltd. The paper was stimulated primarily by recent work by Lee J. Cronbach. Discussions with J. Ward Keesling and Richard J. Shavelson were also instrumental in clarifying the issues addressed. The confusion that remains is solely the authors' responsibility.

[2] The paper by Rock, Baird, and Linn (1972) was cited and discussed in the first author's earlier treatments of multilevel analytical methods (Burstein, 1975, 1976b; Burstein and Knapp, 1975; Burstein and Smith, 1976). Their methodology is not considered here because it is heuristic in nature and does not provide estimates of effects that are interpretable in a regression-based framework. At the same time however, the heuristic algorithm embodied in Rock, Baird, and Linn's generation of clusters for their discriminant analyses is closely akin to the steps carried out in the slope-intercept analyses considered here.

[3] In our computer simulation, we did not generate estimates of the β_c as individual-level effects were of secondary importance to our study. This is an oversight that should be rectified and one that should help clarify the potency of the Keesling-Wiley method in the presence of heterogeneous within-class slopes.

[4] The slope-intercept method described here represents a modification of the procedures proposed earlier (Burstein, 1976b). We have dropped, for the moment, the idea of treating the standard error of estimate as an additional outcome variable and have formed α_c to reduce the strong negative correlation between α_c and β_c across classrooms. The transformation means that

$$\hat{\alpha}_c = \bar{Y}_c .$$

[5] The parameter estimates found in the case of random variation of within-class regression of Y on X provide an indication of the sampling variability due to choice of a particular set of normal deviates. Thus, estimates under other conditions can be expected to differ from the population parameters by an amount due the sampling variability plus by whatever biases are induced by the systematic teacher effects on variation in the within-class slopes. So, e.g., in Table 5 b_{YT} isunderestimated by .4 due to the teacher effects on slope of -.8 in addition to the .07 downward adjustment attributable to sampling variability.

[6] A more accurate specification of the stages in the Keesling-Wiley analysis would include family background in the pooled within-class analysis [equation (8)] and school and quality in the between-class analysis [equation (9)]. For conditions simulated here, the failure to include family background in the within-class model is inconsequential and the failure to include school quality in the between-class model is a detectable model misspecification which is demonstrated by the data from b_{YS} = .2 conditions in Table 8.

Averch, H., Carroll, S.J., Donaldson, T.S., Kiesling, H.J. and Pincus, J. How Effective is Schooling?, The Rand Corporation, R-956-PCSF/RC, 1972.

Berliner, D.C., et al. Proposal for Phase IIIB of the Beginning Teacher Evaluation Study, San Francisco: Far West Laboratory for Educational Research and Development, 1976.

Brown, B.W. and Saks, D.H. "Proper Data Aggregation for Economic Analysis of School Effectiveness." Public Data Use, 1975, 3: 13-18.

Burstein, L. "The Choice of Unit of Analysis in the Investigation of School Effects: IEA in New Zealand. New Zealand Journal of Educational Studies, 1976, 2: 11-24.

Burstein, L. "Assessing the Differences of Between-groups and Individual Regression Coefficients." Paper presented at the American Educational Research Association Annual Meeting, San Francisco, April, 1976.

Burstein, L. "Data Aggregation in Educational Research: Applications." Paper presented at the Annual Meeting of the American Educational Research Association Annual Meeting, Washington, D.C., April, 1975.

Burstein, L. and Knapp, T.R. "The Unit of Analysis in Educational Research." Technical Report No. 4, Consortium on Methodology for Aggregating Data in Educational Research. Milwaukee, Wisc.: Vasquez Associates, Ltd., July, 1975.

Burstein, L. and Smith, I.D. "Choosing the Appropriate Unit for Investigating School Effects." Australian Journal of Education, 1976 (in press).

Cline, M.G. et al. "Education as Experimentation: Evaluation of the Follow Through Planned Vairation Model." Cambridge, Mass.: Abt Associates, 1974, Vols. 1A, 1B (Ed 094890, 094891).

Cronbach, L.J. "Research on Classrooms and Schools: Formulation of Questions, Design and Analysis." Occasional Papers of the Stanford Evaluation Consortium, July 1976.

Cronbach, L.J. and Webb, N. "Between-class and Within-class Effects in a Reported Aptitude Treatment Interaction: Reanalysis of a Study by G. L. Anderson." Journal of Educational Psychology, 1975, 67: 717-727.

Glass, G.V. and Stanley, J.C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Glendening, L. "The Effects of Correlated Units of Analysis: Choosing the Appropriate Unit." Paper presented at the AnnualMeeting of the American Educational Research Association, San Francisco, April, 1976.

Haney, W. "Units of Analysis Issues in the Evaluation of Project Follow Through." Unpublished Report, Cambridge, Mass.: Huron Institute, 1974.

Hannan, M.T., Freeman, J., and Meyer, J.W. "Specification of Models of Organizational Effectiveness." American Sociological Review, 1976, 41: 136-143.

Keesling, J.W. and Wiley, D.E. "Regression Models of Hierarchical Data." Paper presented at the Annual Meeting of the Psychometric Society, Stanford, Calif., 1974.

Lohnes, P.R. "Statistical Descriptors of School Classes." American Educational Research Journal, 1972, 9: 547-556.

Peckham, P.D., Glass, G.V. and Hopkins, K.D. "The Experimental Unit in Statistical Analysis." Journal of Special Education, 1969, 3: 337-349.

Ray, H.W. "Final Report on the Office of Economic Opportunity Experiment in Educational Performance Contracting." Unpublished report, Battelle Laboratories, Columbus, Ohio, 1972.

Rock, D.A., Baird, L.L. and Linn, R.L. "Interaction between College Effects and Students'Aptitudes." American Educational Research Journal, 1972, 10: 149-161.

Scheuch, E.K. Cross-national comparisons using aggregate data: Some substantive and methodological problems. In R.L. Merritt and S. Rokkan (Eds.) Comparing Nations: The Use of Quantitative Data in Cross-national Research, New Haven: Yale University Press, 1966, pp. 131-167.

Wittrock, M.D. and Wiley, D.E. The Evaluation of Instruction: Issues and Problems. New York: Holt, Rinehart & Winston, 1970.

ON CERTAIN SIMILARITIES IN THE ESTIMATION OF MULTI-WAVE PANELS AND MULTI-LEVEL CROSS-SECTIONS

Michael T. Hannan
Stanford University

and

Alice A. Young
Carnegie-Mellon University

Prepared for the Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October 23-24, 1976. This work was supported by NIE Research Contract C-75-0123.

We are struck with similarities in the analysis problems that arise in analysis of multi-wave panel data and multi-level cross-sectional data. These similarities have not been explored in depth and work in each area tends to procede independently of work in the other. In this paper, we begin to explore these similarities. In particular, we argue that the application of the logic recently developed for multi-wave panels to multi-level analysis problems will clarify certain issues that pervade current work on educational methodology.

Most educational systems are hierarchically arranged (in the sense of a partial ordering). This aspect of structure is most visible in primary and secondary education organization. Students are grouped into classrooms within grade levels, within schools, within school districts, within state educational administrations. The methodological challenge is to employ the multiple levels of analysis in modeling the impacts of organizational structure on student careers and outcomes.

Consider the problem of modeling the effects of school structure on student achievement. Suppose we are interested in the effects of some property of school structure (e.g. centralization of curriculum student teacher ratio, etc.) on achievement. There is systematic sorting of families into school districts that produces a correlation of individual student attributes (e.g. ability, socioeconomic status, educational aspirations, etc.) with school characteristics. Therefore, an adequate description of the achievement process must contain both student characteristics and school characteristics.

The practical problem facing the researcher is as follows. He or she may analyze either individual level data (e. g. regressing individual achievement on student characteristics and school characteristics) or he may analyze school level averages (e.g. regress average achievement on average individual characteristics and school properties).[1] We have shown (Hannan et al. 1976) that in most realistic cases (i.e. when the models are not perfectly specified) results of the two different analysis will be quite dissimilar. This finding adds considerable importance to the choice between models.

Some investigators argue that when the question is posed at the school level, the school level regression (or, more correctly, the between-school regression) is most appropriate. A forceful statement of this position is made by Bidwell and Kasarda (1975, 1976). Others (Hauser 1971; Hannan et al. 1976) point out that researchers seldom adequately specifiy school-level process that are anything more than the processes sum of individual level processes. The causal arguments concern impacts on individual students which are composed (by a very simple function, that of addition) to school level outcomes. Consequently, the choice of level is open to question.

To date, the discussion of the issues involved does not appear to have been resolved. Each group holds to its initial position. Therefore, it is valuable to search for a modeling strategy that turns the issue of taste into an empirical problem. We wish to propse that the framework we have discussed elsewhere for multi-wave panel analysis, the "pooling of cross-sections and time series model" (Hannan and Young 1974; 1976) provides a means for doing just that.

It will simplify exposition to begin first with the framework as it applies to panel analysis and then draw the analogues to hierarchical structure analysis. Suppose we are analyzing the effect of X on Y in a panel design in which we have multiple waves, say T of them, on N individuals. Then the simplest model we can employ is

$$Y_{it} = \alpha + \beta X_{it} + u_{it}. \qquad (i = 1,..., N; t = 1,...,T) \qquad (1)$$

where all observations are pooled into a single model.

Suppose a researcher analyzed such a data structure by means of ordinary least squares (OLS) regression. We would almost certainly object that the observations are not independent; the responses of the same individual at different periods are presumably all influenced by his/her unobserved but relatively constant characteristics (e.g. personality, history). OLS depends on the assumption that observations are independently distributed; efficiency is reduced by correlated observations and tests of significance and confidence intervals calculated by OLS formulas are biased (in the "liberal" direction) in such cases.

When dealing with data structures in which left out variables (summarized as disturbances) are correlated, we usually turn to generalized least squares (GLS) (Goldberger 1973).[2] As the name suggests this method is a generalization of OLS that has OLS as a special case when the observations are indeed independent (in the sense that the disturbances are uncorrelated). To see how it handles the autocorrelation problem for panel analysis, we must state our model more explicitly. It will help to write the model in matrix form:

$$\begin{aligned} &\underline{Y}_t = \underline{X}_t\,\underline{\gamma} + \underline{u}_t \\ &Y = (Y_{11},\ldots,Y_{it},\ldots,Y_{NT})' \\ &X = (X_{11},\ldots,X_{it},\ldots,X_{NT})' \qquad (2) \\ &u = (u_{11},\ldots,u_{it},\ldots,u_{NT})' \\ &\gamma = (\alpha,\ \beta)' \end{aligned}$$

and each element of $\underline{u}$ say u_{it} contains a systematic component μ_i and a random disturbance ν_{it}:

$$u_{it} = \mu_i + \nu_{it} \quad .$$

Further,

$$\underline{E}\underline{uu}' = \sigma^2 \begin{bmatrix} \underline{A} & 0 & . & . & . & 0 \\ 0 & \underline{A} & & & & 0 \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ 0 & 0 & . & . & . & \underline{A} \end{bmatrix} \text{, an (NT x NT) matrix ,}$$

where

$$\sigma^2 = \sigma_\mu^2 + \sigma_\nu^2 \; , \quad \rho = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_\nu^2) \text{ , and } \underline{A} = \begin{bmatrix} 1 & \rho & . & . & . & \rho \\ \rho & 1 & & & & \rho \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ \rho & \rho & . & . & . & 1 \end{bmatrix} \text{, a (T x T) matrix.}$$

The mechanism that produces autocorrelation is quite simple; it states that there are omitted variables that are constant for each individual and affect the dependent variable in the same way every period. Consequently all disturbances for each individual are correlated at the same level, call this ρ. Moreover to simplify the model we have assumed that every individual is characterized by the same correlational structure in the distrubances. This is the model that Nerlove (1971) (see also Hannan and Young 1976) applied to the panel analysis problem.

The GLS strategy is to find a matrix Ω that has the form that:

$$Y_t \, \Omega^{\frac{1}{2}} = X_t \, \Omega^{-\frac{1}{2}} \beta + u \, \Omega^{\frac{1}{2}} \tag{3}$$

and

$$\text{Eu } \Omega \text{ u}' = \sigma^2 \begin{bmatrix} 1 & 0 & . & . & . & . & 0 \\ 0 & 1 & . & . & . & . & 0 \\ . & & & & & & \\ . & & & & & & \\ . & & & & & & \\ 0 & . & . & . & . & . & 1 \end{bmatrix}$$

That is, the matrix has the form that once the model is transformed, the resulting disturbances have the form under which OLS is optional. Then, in a sense, one applies OLS to the transformed data. In fact the two steps are done simultaneously. The GLS estimator is defined as

$$\underline{\hat{\beta}}_{GLS} = (\underline{Q}_t'\underline{\Omega}^{-1}\underline{Q}_t)^{-1}\underline{Q}_t'\underline{\Omega}^{-1}\underline{y}_t, \qquad (4)$$

where

$$\underline{\Omega}^{-1} = \begin{bmatrix} \underline{A}^{-1} & 0 & . & . & . & 0 \\ 0 & A^{-1} & . & . & . & 0 \\ . & & . & & & \\ . & & & . & & \\ . & & & & . & \\ 0 & 0 & . & . & . & \underline{A}^{-1} \end{bmatrix},$$

and (cf. Hannan and Young, 1974)

$$\underline{A}^{-1} = (1/\eta)(\underline{I}_T - \underline{\ell\ell}'/T) + (1/\xi)(\underline{\ell\ell}'/T),$$

where $\eta = (1 - \rho)$, $\xi = (1 - \rho) + T\rho$, and $\underline{\ell}$ is a $(T \times 1)$ vector of ones.

The form of the GLS transformation (4) can be intuitively motivated as follows. The peculiar feature of pooled models is the use of both cross-sectional (between-unit) and longitudinal (within-unit) variation to estimate causal parameters. The richness of the data presents an implicit choice: how to weight one type of variation relative to another. Generalized least squares uses ρ to weight the two types of information. To see this, consider the case where $\rho = 0$. Then $\underline{A}^{-\frac{1}{2}} = \underline{I}_T$ and observations are transformed in (3) by an identity transformation. GLS reduces to OLS where cross-sectional and time series variation are weighted proportionately to N and T (see Maddala, 1971). At the other extreme, when $\rho = 1$, $\underline{A}^{-1} = \underline{\ell\ell}'/T^2$. It is easy to show that this transformation averages observations over time for each unit. The result is a regression on grouped observations where all of the weight is placed on cross-sectional variation. In cases where ρ takes on a value $0 \leq \rho \leq 1$, GLS weights time series variation inversely to ρ. Such a weighting is appropriate since ρ measures "redundancy" in the time series. The more redundancy, the lower the weight attached to longtitudinal variation.

The gain in efficiency of GLS relative to OLS arises because GLS uses a weighting based on population or sample information rather than an arbitrary weighting (cf. Goldberger, 1973). Such a gain can be realized whenever ρ is known a priori or can be estimated from sample data in a consistent (or unbiased) manner. When GLS is applied using the simple form of estimation of $\hat{\rho}$ proposed by Nerlove, GLS has very good small sample properties (Nerlove 1971; Hannan and Young 1974, 1976).[3]

Now that we have stated the technical details of the model we can return to the substantive problem. That concerns the choice among two models for the achievement study example. The first is an individual level analysis of all students that treats each student as an independent observation. The second alternative is a between-school regression that ignores within-school variations. The analogy between this situation and the multi-wave panel problem is immediate. Schools play the role of individuals in the panel design and students play the role of observations on individuals; that is individuals are organized hierarchically within schools in the same manner that individual outcomes at several points in time in the panel are organized within individuals. Outcomes to students in the same school are likely to be nonindependent for the same reason that responses by the same individual over time are nonindependent: the model will almost certainly ignore some causal variables that are stable for schools (or individuals) over pupils (or time). For example, students in the same school are exposed to some extent to a common school culture, to common facilities, have common histories, etc.

Given this close similarity there appears to be much to be gained from reformulating the hierarchical structure analysis problem in GLS terms. If we formulate the model in terms of a state space defined in terms of the lowest level units (e.g., students) and include components to variance for each higher level units, we see that in most cases the optimal statistical estimators of organizational effects will be neither the individual level model nor the between-school models discussed above. These models are appropriate only in extreme cases, namely when individual outcomes are completely uncorrelated or perfectly correlated. In all other cases, the appropriate model will fall between these two and will depend on the actual levels of intercorrelation in the sample.

To this point we have assumed proper model specification. It is time to consider the way in which the analysis presented above must be modified when the models being estimated do not include all of the relevant causal variables. Elsewhere, we have considered the divergence of grouped and ungrouped estimators (between and total regressions) when the omitted variables vary both between and within groups (Hannan et al., 1975; Hannan and Young 1976 b). The results for such cases are relatively straight forward: for most realistic cases, grouping magnifies the specification bias of the ungrouped estimators. The type of generalized least squares analysis discussed here does not handle this type of problem.

For present purposes, it is more useful to consider a model that omits causal factors that vary only between groups (e.g., school climate, professional staff size per pupil) and others that vary within, but not between groups (e.g., rank in class). For example, consider:

$$Y_{it} = \beta X_{it} + (\mu_i + \tau_t + v_{it}) \qquad (5)$$

where both μ and τ, the school variable and the within-school variable respectively, are both uncorrelated with V but are correlated with X.

Now consider three alternative OLS estimators of β:

$$\hat{\beta} = \sum_i \sum_t Y_{it} X_{it} / \sum_i \sum_t X_{it}^2 \qquad \text{(total regression)}$$

$$\hat{\beta}(b) = \sum_i Y_{i\cdot} X_{i\cdot} / \sum_i X_{i\cdot}^2 \qquad \text{(between regression)}$$

and

$$\hat{\beta}(w) = \sum_i \sum_t (Y_{it} - Y_{i\cdot})(X_{it} - X_{i\cdot}) / \sum_i \sum_t (X_{it} - X_{i\cdot})^2 \qquad \text{(within regression)}$$

All three estimators are biased; however, each has a different bias. More precisely:

$$\text{plim}\,\hat{\beta} = \beta + A + B$$
$$\text{plim}\,\hat{\beta}(b) = \beta + A$$
$$\text{plim}\,\hat{\beta}(w) = \beta + \quad B$$

where A is the sum of the sample regressions of the μ on X and B is the sum of the sample regressions of the V on X. Depending on the magnitudes and signs of A and B, any one of the three estimators may be superior to the others. For example, if A + B $\hat{=}$ 0, the total regression gives good estimates of B on average while neither of the others do.

In such cases then, choice among estimators affects both bias and efficiency. Suppose in the example just considered that A + B and, consequently, all three OLS estimators are biased and inconsistent. If one can determine a priori the nature of the confounding between and within variables, one can form consistent GLS estimators by the procedures discussed by Nerlove (1971). In this case, one must estimate variance components that parallel the individual-specific and time period specific effects considered in multi-wave panels.[4] If the variances and covariances of the omitted variables have the simple form assumed by Nerlove, then modified GLS gives good small sample estimators of β (Young 1976). In this case again, the most desirable estimator is neither the total, the between, nor within OLS estimators. Instead, it is one that uses sample information to weight between and within information in an optimal way.

Of course, the situation in empirical research is more complex. The nature of the omitted variables are never known precisely and they are seldom as well behaved as we have assumed here. Nonetheless, we believe that the approach sketched here is an useful starting point for rethinking the analysis of multi-level data structures.

FOOTNOTES

[1] There is, of course, a third possibility: a pooled within-school analysis. However, the total, within and between analyses, are not independent. Thus, we can choose any two of the three for the present discussion. Most of the literature that has considered these problems deals with between and total relations. One could as well consider between and within without altering anything in the argument (though model misspecification changes this as we indicate below).

[2] If one considers the omitted individual effects to be fixed effects, one handles the non-independence by including dummy variables for each individual (or alternatively expressing all observations are deviations from individual means). Clearly, this representation is algebraically identical to the within-individual regression.

[3] When the process generating the disturbances is as stated here, MGLS is also more efficient that the procedure of including dummy variables for each individual, though the two estimators converge asymptotically. The efficiency gain in small samples is presumably due to the fact that the MGLS procedure utilizes far fewer degrees of freedom.

[4] One can as well estimated a fixed-effects model for this case. Then one introduces dummy variables for schools and class ranks, say, and performs OLS (or equivalently, one can conduct a pooled within-school, within-rank OLS analysis). However, Young (1976) finds that when the omitted variables have the simple form considered here, this procedure has small sample properties that are much inferior to modified generalized least squares. Nonetheless, the two procedures are asymptotically equivalent.

REFERENCES

Bidwell, Charles E. and John D. Kasarda
1975 "School District Organization and Student Achievement", American Sociological Review, 40: 55-70.

Cronbach, Lee J.
1976 Educational Effects at the Individual and Group Levels: The Crucial Importance of Hierarchical Interpretation. Final report to the Spencer Foundation.

Goldberger, Arthur S.
1973 "Efficient Estimation in Over-Identified Models: An Interpretive Overview", in A. S. Goldberger and O. D. Duncan (eds.) Structural Equation Models in the Social Sciences, New York, Seminar Press.

Hannan, Michael T., John Freeman and John W. Meyer
1976 "Specification of Models of Organizational Effectiveness: Comment on Bidwell and Kasarda", American Sociological Review, 41: 136-143.

Hannan, Michael T. and Alice A. Young
1976 a "Estimation in Multi-Wave Panel Models: Results on Pooling Cross-Sections and Time Series" in D. Heise (ed.) Sociological Methodology, 1977, San Francisco, Jossey-Bass (in press).

1976 b "Small Sample Results on Estimation from Grouped Observations".

Maddala, G. S.
1971 "The Use of Variance Components Models in Pooling Cross-Section and Time Series Data", Econometrica, 39: 341-358.

Nerlove, Marc
1971 "Further Evidence on the Estimation of Dynamic Economic Relations From a Time Series of Cross-Sections", Econometrica, 39: 359-382.

Young, Alice A.
1976 "Organizational Growth: A Methodological Analysis," unpublished dissertation, Stanford University.

COMPONENTS OF VARIANCE MODELS IN MULTILEVEL VARIANCE DATA

J. W. KEESLING

This paper begins with the premise that an interesting class of analyses involves settings in which pupils are sampled within larger groupings (school, classroom) and where both the pupil dimension and the grouping dimension are appropriately considered to be random factors. I propose to present a model (developed by Schmidt, 1969) which generalizes the standard univariate components of variance model to the multivariate case and will permit us to look at comparisons of between groups and within groups componen of variance and covariance.

Univariate Random Effects Model

(1) $$y_{ij} = \mu + a_i + e_{ij} \qquad i=1,\ldots,k \qquad j=1,\ldots,n$$

$$e_{ij} \sim NID(0, \sigma^2) \qquad Cov(a_i, e_{ij}) = 0$$

$$a_i \sim NID(0, \sigma_a^2) \qquad \sigma_y^2 = \sigma_a^2 + \sigma^2$$

Substantively, y_{ij} may be thought of as the score of the j^{th} subject in the i^{th} group on a test. μ is an over-all mean, a_i is the group effect and e_{ij} is the individual pupil effect. The inferential problem concerns estimating σ_a^2 and σ^2. Usually Analysis of Variance is used to obtain the needed estimates. It can be shown that:

$$E[MS_{Within}] = \sigma^2 \qquad \hat{\sigma}^2 = MS_W$$

$$E[MS_{Between}] = \sigma^2 + n\sigma_a^2 \qquad \hat{\sigma}_a^2 = \frac{1}{n}(MS_B - MS_W)$$

Note, however, that in this analysis there is a non-zero probability of obtaining $\hat{\sigma}_a^2 < 0$. Tiao and Tan (1965, 1966) have approached the problem from another viewpoint. They regard the setting as one in which k groups of n observations are drawn independently from a n-dimensional multivariate normal distribution with mean $\mu \underline{1}_n$ and covariance matrix:

$$\Sigma = \sigma^2 I_n + \sigma_a^2 \underline{1}_n \underline{1}_n' ; \quad \underline{1}_n \text{ is an } n \times 1 \text{ vector of ones.}$$

The implication is that there are not kn independent observations in the data set - only k. The y_{ij} covary within a given group i:

$$Cov(y_{ij}, y_{ij'}) = \sigma_a^2 \qquad Cov(y_{ij}, y_{i'j}) = 0$$

Multivariate Random Effects Model

The multivariate generalization of (1) is:

(2) $\underline{y}_{ij} = \underline{\mu} + \underline{a}_i + \underline{e}_{ij}$ where each vector is p x 1.
$i = 1, \ldots, k \quad j = 1, \ldots, n$

We may also write: $\Sigma_y = \Sigma_a + \Sigma$

And again there is an Analysis of Variance solution for this model which may be summarized as:

$$S = \sum_{i=1}^{k} \sum_{j=1}^{n} (\underline{y}_{ij} - \underline{\bar{y}}_{i.})(\underline{y}_{ij} - \underline{\bar{y}}_{i.})'$$

$$S_a = n \sum_{i=1}^{k} (\underline{\bar{y}}_{i.} - \underline{\bar{y}}_{..})(\underline{\bar{y}}_{i.} - \underline{\bar{y}}_{..})'$$

$$E\left[\frac{1}{k(n-1)} S\right] = \Sigma ; \qquad \hat{\Sigma} = \frac{1}{k(n-1)} S$$

$$E\left[\frac{1}{k-1} S_a\right] = \Sigma + n\Sigma_a ; \quad \hat{\Sigma}_a = \frac{1}{n}\left[\frac{1}{k-1} S_a - \frac{1}{k(n-1)} S\right].$$

This solution suffers from the multivariate analog of negative variance estimates: There is a non-zero probability that $|\hat{\Sigma}_a| \leq 0.$

Schmidt (1969) adapted the notion developed by Tiao and Tan (1965, 1966) to the multivariate case by supposing that k groups of n·p-dimensioned observations are independently drawn from an n·p-dimensional multivariate normal distribution with mean $\underline{1} \otimes \underline{\mu}$ and coveriance matrix:

$$\Sigma_{np} = \underline{1}\,\underline{1}' \otimes \Sigma_a + I \otimes \Sigma.$$

For a single sample (one of the k such samples) we may write:

$$(3) \qquad \underset{np\times 1}{\underline{y}} = \underset{n\times 1}{\underline{1}} \otimes \underset{p\times 1}{\underline{\mu}} + \underset{n\times 1}{\underline{1}} \otimes \underset{p\times 1}{\underline{a}} + \underset{np\times 1}{e}$$

which implies that

$$Cov(\underline{y}_{ij}, \underline{y}_{ij'}) = \Sigma_a \text{ for a given } i.$$

The density function of $\underline{y}$ may be written as:

$$f(y) = (2\pi)^{-np/2} |\Sigma_{np}|^{-1/2} exp\{-\tfrac{1}{2}[(\underline{y} - \underline{1} \otimes \underline{\mu})' \Sigma_{np}^{-1} (\underline{y} - \underline{1} \otimes \underline{\mu})]\}.$$

The log of the likelihood function has the form:

$$k(1-n)/2 \log|\Sigma| - (k/2) \log(|\Sigma + n\Sigma_a|) - (kn/2)\, tr(\Sigma^{-1}S) - (k/2)\, tr[(\Sigma + n\Sigma_a)$$

From which the maximum likelihood estimators

$$\hat{\Sigma} = \frac{n}{n-1} S \text{ and } \hat{\Sigma}_a = \frac{1}{k}\left(S_a - \frac{n}{n-1} S\right)$$ are obtained.

Schmidt refers to these as unrestricted estimators in the sense that the elements of $\hat{\Sigma}$ and $\hat{\Sigma}_a$ may take on any values in the appropriate parameter space. A further parameterization yields a familiar model for each of Σ and Σ_a:

$$(4) \qquad \Sigma_a = \Lambda_a \Phi_a \Lambda_a' + \Psi_a$$

and

$$\Sigma = \Lambda \Phi \Lambda' + \Psi$$

Models of this form have been investigated by Jöreskog (1970 being only one example) and Wiley, Schmidt and Bramble (1973) among others. Here the interest lies in the possibility of specifying different relational models for the between components of covariance matrix and the within components of covariance matrix. Schmidt (1969) also considers a case where the Λ matrix is common across Σ_a and Σ.

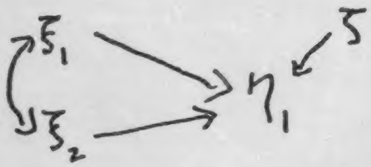
Jöreskog (1970) has shown that models of the form $\Sigma = \Lambda \Phi \Lambda' + \Psi$ may be explicitly parameterized to represent a wide variety of factor analytic and path analytic models. This implies that Schmidt's formulation should give us an analytic technique which could be used, as Cronbach (1976) suggests, to formulate quite disparate models for between-group and within-group structural relationships.

As an example, suppose that we could assess achievement of youngsters prior to the onset of instruction and again after instruction. Suppose that we also had a measure of the instructional time devoted to each student. We could construct the simple model represented as a path diagram below:

ξ_1 = pretest score
ξ_2 = instructional time
η_1 = posttest score

$$\eta_1 = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \zeta$$

where ζ is a residual.



A parameterization of Σ could be:

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \gamma_1 & \gamma_2 & 1 \end{bmatrix} \quad \Phi = \begin{bmatrix} \sigma^2_{\xi_1} & & \text{Sym} \\ \sigma_{\xi_1 \xi_2} & \sigma^2_{\xi_2} & \\ 0 & 0 & \sigma^2_{\zeta} \end{bmatrix} \quad \Psi = [0]$$

The corresponding parameterization of Σ_a could be:

$$\Lambda_a = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \gamma_{1a} & \gamma_{2a} & 1 \end{bmatrix} \quad \Phi_a = \begin{bmatrix} \sigma^2_{\xi_{1a}} & & \\ \sigma_{\xi_{1a} \xi_{2a}} & \sigma^2_{\xi_{2a}} & \\ 0 & 0 & \sigma^2_{\zeta a} \end{bmatrix} \quad \Psi_a = [0]$$

One particular hypothesis which would be of interest in this setting would be $\gamma_2 = \gamma_{2a}$; that is, the structural coefficient relating posttest score to instructional time would be the same for pupils within groups as it is across groups. By comparing the fit of the model where $\hat{\gamma}_2$ is constrained to equal $\hat{\gamma}_{2a}$ to the fit of the model in which there is no such constraint, one can obtain a test of the hypothesis specified above.

The analytic technique due to Schmidt (1969) should prove useful in studies comparing structural hypotheses concerning between group and within group components of covariance. Further work will be necessary to extend the model to additional hierarchical levels and to provide for parameterization of a more comprehensive set of models - e.g., the full set described by Jöreskog (1970).

REFERENCES

Cronbach, L. J., "Research on Classrooms and Schools: Formulation of Questions, Design and Analysis", Stanford Evaluation Consortium, 1976.

Jöreskog, K. G., "A General Method for Analysis of Covariance Structures", Biometrika, 1970, 57: 239-251.

Schmidt, W. H., "Covariance Structure Analysis of the Multivariate Random Effects Model", unpublished Ph. D. dissertation, Department of Education, University of Chicago, June, 1969.

Tiao, G. C. and Box, G. E. B.,"Bayesian Analysis of a Three-Component Hierarchical Design Model", Biometrika, 1967, 54: 109-125.

Tiao, G. C. and Tan, W. Y., "Bayesian Analysis of Random-Effect Models in The Analysis of Variance", I. Posterior Distributions of Variance Components", Biometrika, 1965, 52: 37-53.

Tiao, G. C. and Tan, W. Y., "Bayesian Analysis of Random Effect Models in the Analysis of Variance, II, Effect of Auto-Correlated Errors", Biometrika, 1966, 53: 477-495.

Wiley, D. E., Schmidt, W. H. and Bramble, W. J., "Studies of a Class of Covariance Structure Models", Journal of the American Statistical Association, 1973, 68: 317-323.

COMMENTS PREPARED FOR CONFERENCE ON METHODOLOGY FOR AGGREGATING DATA IN EDUCATIONAL RESEARCH

(STANFORD UNIVERSITY, OCTOBER 23-24, 1976)

Charles R. Henderson, Jr.

Cornell University

Based on the papers I have had the opportunity to read in advance of the conference are a few comments:

(1) It is essential correctly and fully to specify the model in any given analysis. (There should be some model appropriate to attempt to answer the questions of interest.) This means including all relevant classification factors and covariates in the model and, for example, making the proper assumptions regarding which factors are fixed and which are random. Then there will be some best way to analyze the model (best linear unbiased estimation, exact hypothesis tests, etc.) and one should attempt such an analysis. Of course, BLUE and exact tests cannot always be achieved, in particular in mixed models with covariates and/or unequal smallest subclass numbers in the case of unknown variance components (the usual case). Still, one should use the best methods possible. The fixed vs. random distinction is always crucial (except in the special case where the function of interest is estimable in a model in which the random factors are treated as fixed for purposes of examining estimability), and is especially so in models which include higher level covariate regressions (where all observations on the same level of a factor or interaction have the same covariate value).

(2) Analysis of mixed models with covariates and/or unequal numbers requires general linear model methods, and few useful generalizations are possible from the fixed model. While exact hypothesis tests are not, in general, possible when variance components are unknown, a minimum requirement for a test would seem to be that it is unbiased in the sense that the expectations of the numerator and denominator of the test statistic are equal under the null hypothesis. To obtain such tests requires estimation of the function of interest, computation of the variance of the estimator which involves the unknown variance components. The variance components need to be estimated and substituted in the variance of the estimator. These general methods and the above points are discussed in the three articles I am distributing at the conference.

(3) In general, methods which are essentially computing algorithms based on some special case, will not give desirable (e.g., small sampling variance of estimators, unbiased hypothesis tests) results. These methods appear to come from a desire to use known simple analysis of variance models and/or canned computer programs. Forcing real models to fit these models frequently leads to treating random factors as fixed or ignoring factors completely.

Therefore, for example, the emphasis should not be in terms of what is the "unit of analysis," but rather on including all factors in the model. Working with a model written for computational purposes by taking averages on original dependent variable and covariate values over levels of a factor in order to, in effect, include that factor in the model does not work in general. Likewise, doing covariance analysis by estimating the covariate regression coefficients in the fixed model, adjusting the dependent variables for the covariates, and then carrying out analysis of variance on the adjusted model, does not give unbiased tests unless generalized least squares is used since even if the original error variance was $\sigma^2 \underset{\sim}{I}$, the error variance in the adjusted model has general form.

(4) There seems to be a need to argue for classrooms as the "unit of analysis" in the case when units are not independent (certainly the usual case) which is to say that classrooms should be a factor in the model as a random factor. This concept is basic, and applies equally to all other relevant factors. (Including random factors, with their distribution assumptions, deals with the problem of "correlated units"; including fixed factors removes bias.)

(5) Models which include higher level covariates appear to be in use. (Here we do not mean models where data are not available on the individual observations and aggregate data must be used on both dependent variable and covariates, or where averages are taken on a factor in an attempt to include the factor in the model.) Such models are complicated, and are discussed in the articles. Estimability considerations lead to the conclusions that higher level covariates should only be defined on random factors or interactions, not fixed, and that such covariates should not even be estimated in a model which treats random factors as fixed.

REFERENCES

Henderson, C.R., Jr. [1976]. Statistical analysis of a complex model involving fixed and random factors, unequal subclass numbers, and higher level covariates. Submitted for publication.

Henderson, C.R., Jr. and Henderson, C.R. [1976]. Analysis of covariance in mixed models with unequal subclass numbers. Submitted for publication.

Henderson, C.R., and Henderson, C.R., Jr.[1976]. Estimation, prediction, and hypothesis testing in the mixed linear model with unequal subclass numbers. Submitted for publication.

Structural Equation Models of Contextual Effects*

Duane F. Alwin

Indiana University

*Paper prepared for presentation at a Conference on Methodology for Aggregating Data in Educational Research, sponsored by a grant from the National Institute of Education (NIE-C-74-0123), co-organized by Leigh Burstein and Michael T. Hannan. Stanford University, Stanford, California. October 23-24, 1976.

Structural Equation Models of Contextual Effects

Preface

This paper combines the results of two recent analyses of school context effects. The first (Alwin, 1976) is a didactic presentation of the structural equation form of the analysis of covariance model proposed for assessing school context effects (Hauser, 1968, 1969, 1971). This discussion emphasizes the relationship of the analysis of covariance and the use of compositional variables in structural equation models for describing contextual effects. The second (Alwin and Otto, 1976) is an empirical analysis of contextual effects in a substantive domain characterized by contextual effects of opposite signs. This analysis discusses the specification problems introduced by so called "frogpond" phenomena, and assesses their presence in sample data using a structural equation model.

Introduction

Despite the conclusions of the Equality of Educational Opportunity report (Coleman et al., 1966) and major reanalyses of the EEO data (Mosteller and Moynihan, 1972), that the distribution of resources among U. S. public schools is more even than was once believed, sociologists continue to nurture an interest in assessing the impacts of school differences on students' aspirations and achievements. This is indicated by a prevailing view that because residential segregation occurs largely on the basis of socioeconomic criteria (see Wilson, 1959), differences in the socioeconomic context of the school may be important in shaping the academic performances, self conceptions, aspirations and achievements of students. Whether school contexts are differentiated in terms of socioeconomic factors (Rogoff, 1961; Turner, 1964), peer associations (Campbell and Alexander, 1965), ability contexts (Meyer, 1970; Alexander and Eckland, 1975), educational climates (McDill et al., 1969), or unmeasured contextual factors, sociologists who have studied the problem have, with a few exceptions, been optimistic that the school context would prove to be an important determinant of individual outcomes.[1]

It is not always recognized in the study of school effects that individual-level variables intended to represent processes which operate within groups and group-level variables intended to represent group effects are not easily "unconfounded" at the level of measurement. The estimation of either within-group or between-group effects presupposes the proper specification of the theoretical process at both levels. Therefore, unless the within-group process can be adequately specified, any inference about group effects is risky unless one can be assured that omitted individual variables and group-level variables are unrelated. The recognition of this type of problem led Hauser (1970a) to identify erroneous inferences about school effects where the within-school process is not adequately specified as a class of "contextual fallacy."[2] Hauser's contextual fallacy is indifferent to the particular manner in which the group-level process has been represented, and one may be vulnerable to committing this type of inferential error whether he uses compositional variables, the analysis of covariance or some other method of measuring group effects (see Hannan et al. (1976) comment on Bidwell and Kasarda (1975)). The erroneous inference about within-group effects where the between-group process is not properly specified is possible as well, and the concern with this type of specification bias is largely responsible for the search for school-level variables which explain variation in individual-level outcomes. This section of the paper discusses the implications of the problem of between-group specification in school effects studies.[3] We begin by discussing a general model that states the issue of between-school specification, and then develop a description of two approaches to the problem. We briefly discuss the robustness of the model given the problems of measurement error and selection bias. Finally, we present an empirical analysis using the proposed model.

The Issue of Between-school Specification--the Simple Case

We begin the discussion with a simple structural equation for a dependent variable of interest, Y, stated for some known population.

$$Y_{ij} = a_{YX} + b_{YX} X_{ij} + e_{ij} \qquad (1)$$

The index j varies over schools and the index i varies over individuals within schools. To give substance to this simple model, suppose variable X is father's educational attainment and Y is son's college plans. Equation 1 states that Y is a linear function of X and a random disturbance, e. The minimal set of assumptions for the disturbance term here is that its mean is zero in the population, its variance is constant over levels of X and independent of X, and the values of e are independent of one another (see Johnston, 1972). We also assume perfect measurement in both variables.[4] In this model the effect of X on Y is represented by the parameter b_{YX}, interpreted as the amount Y would change for a given unit shift in X. In our substantive example, b_{YX} is the number of units or some part of a unit change in son's college plans associated with an hypothetically manipulated one unit shift in father's education. Or, it may be thought of as the difference in college plans expected for persons separated by one unit on the scale of father's education.

The assumptions of the model in equation 1 may be wrong. In particular, our interest in the effect of X on Y may extend to the effects operating within and between categories of a third variable, Z, in this case schools. Now we express X's effect on Y in terms of the effects of its within- and between-school components, as follows:

$$Y_{ij} = a'_{YX} + b_{Yx} (X_{ij} - \overline{X}_{.j}) + b_{Y\overline{X}} \overline{X}_{.j} + e'_{ij} \qquad (2)$$

Equation 2 states that Y is a linear function of two uncorrelated variables, a within-school component $(X_{ij} - \overline{X}_{.j})$ and a between-school component $(\overline{X}_{.j})$, plus a random disturbance, e'_{ij}, as in equation 1. The intercept and error term in eq. 2 are not in general the same as those in eq. 1, so we have designated them differently. As above, this model assumes that the error term has a zero population mean, the error variance is constant and independent of X, and the e'_{ij} are mutually unrelated. Again, we assume perfect measurement in both X and Y, and in addition we assume that there is no explicit selection or "grouping" of students by school on the basis of the dependent variable (Hannan and Burstein, 1974). We return to these issues below. (Note that in our subscript notation for the parameters in these equations we refer to the within component as x and to the between component as $\overline{X}$.) The within-school effect is interpreted as the amount Y would change for a unit shift in the deviation of X from the school mean on X, i.e. $X_{ij} - \overline{X}_{.j}$. In this statement of the model, the effect b_{Yx} is assumed to be the same within each school in the population, i.e. the within-school effect is independent of any particular school. The between-school effect, $b_{Y\overline{X}}$, is interpreted as the amount Y would change for a unit shift in the school mean on X independent of variability within groups.[5]

If the model in eq. 1 is the true specification for the relationship between X and Y, then it follows that $b_{Yx} = b_{Y\overline{X}} = b_{YX}$. These are the necessary and sufficient conditions for reducing eq. 2 to eq. 1. However, by entertaining the hypothesis that $b_{Yx} \neq b_{Y\overline{X}}$ we raise the question of between-school specification and the possibility of school effects. Indeed, as the presentation below shows, the possibility of school context effects hinges on whether b_{Yx} and $b_{Y\overline{X}}$ are equal. We must be careful to make explicit the differences implied by within- and between-school processes. When processes affecting the outcomes of schooling occur in relatively uniform fashion within schools, like teachers' or counselors' encouragement to attend college or the effects of curriculum placement (Hauser et al., 1976), we refer to this type of effect as an effect of schooling, not as a school effect. We reserve the latter term for effects which occur by virtue of school-to-school variation in relevant individual-level variables or structural properties of the school. An understanding of this distinction is crucial to properly distinguishing school effects from other effects of schooling processes.

In our substantive example, if b_{Yx} is nonzero, father's education has a nonzero linear effect on son's plans regardless of the particular school attended by the son (a within-school effect). If $b_{Y\overline{X}}$ is nonzero, compositional differences among schools in the education of fathers have a nonzero linear effect on son's plans. A nonzero between-school effect is by itself insufficient evidence of a school effect. As we shall point out in more detail below, the between-school effect may be decomposed into two correlated components, one reflecting the spurious effect of school composition on X, and one reflecting unmeasured factors related to school composition, but representing the "contextual" effect. Both methods described below are aimed at isolating the part of the between-school effect which is the school context effect.

Some Preliminaries

Before describing the two methods of measuring school effects and showing how they are related it is important to establish some relationships which will be helpful for these purposes. First, we should establish that the least squares regression estimates for the parameters in eq. 2 are simply

$$b_{Yx} = \mathrm{Cov}(Y,x) \,/\, \mathrm{Var}\,(x) \qquad (3)$$

$$b_{Y\overline{X}} = \mathrm{Cov}(Y,\overline{X}) \,/\, \mathrm{Var}\,(\overline{X}) \qquad (4)$$

Second, suppose we were to partition the dependent variable Y into its within- and between-school components, as we have already for the independent variable in eq. 2. This results in $Y_{ij} = (Y_{ij} - \overline{Y}_{.j}) + \overline{Y}_{.j}$. Then suppose we were to perform a least squares regression of the within-school portion of Y on the within-school portion of X as follows:

$$(Y_{ij} - \overline{Y}_{.j}) = b_{yx}\,(X_{ij} - \overline{X}_{.j}) + w_{ij} \qquad (5)$$

where w has the properties of disturbances discussed earlier, that it has a zero mean in the population, constant variance, zero correlation with $(X_{ij} - \overline{X}_{.j})$, and its values are independent of one another. Similarly, suppose we were to carry out the regression of the between portion of Y on the between portion of X, as follows:

$$\overline{Y}_{.j} = a_{\overline{YX}} + b_{\overline{YX}}\overline{X}_{.j} + h_{ij} \qquad (6)$$

where h_{ij} is a disturbance, uncorrelated with $\overline{X}_{.j}$, etc.[6] In these regression equations the least squares estimates for the regression coefficients are

$$b_{yx} = \mathrm{Cov}(y,x) \;/\; \mathrm{Var}\ (x) \qquad (7)$$

$$b_{\overline{YX}} = \mathrm{Cov}(\overline{Y},\overline{X}) \;/\; \mathrm{Var}\ (\overline{X}) \qquad (8)$$

Given our assumptions about errors, it is well-known that Cov(Y,x) = Cov(y,x) and Cov(Y,$\overline{X}$) = Cov($\overline{Y}$,$\overline{X}$), and it follows that $b_{Yx} = b_{yx}$ and $b_{Y\overline{X}} = b_{\overline{YX}}$. So, we may identify either of the coefficients b_{Yx} or b_{yx} as the within-school effect in the simple model discussed here, and we may refer to either $b_{Y\overline{X}}$ or $b_{\overline{YX}}$ as the between-school effect (see Schuessler, 1971). We mentioned earlier that the issue of specifying the effect of X on Y in terms of eq. 1 versus eq. 2 depends upon whether $b_{Yx} = b_{Y\overline{X}}$. This can now be restated in terms of b_{yx} and $b_{\overline{YX}}$. If the difference between these parameters is zero, the specification of a between-school process, allowing for school effects, is unnecessary. Both methods we discuss below focus precisely on this difference.

The Analysis of Covariance and School Effects

Using the notation established above we now present a model developed by Hauser (1968, 1971) for the analysis of covariance. This model is as follows:

$$(Y_{ij} - \overline{Y}_{.j}) = b_{yx}\ (X_{ij} = \overline{X}_{.j}) + w_{ij} \qquad (9)$$

$$\overline{Y}_{.j} = b_{yx}\ (\overline{X}_{.j} - \overline{X}_{..}) + v_{j} \qquad (10)$$

Eq. 9 states that the within-school component of Y is a linear function of the within-school component of X and a random disturbance term. The coefficient b_{yx} is the within-school regression coefficient as above in eq. 5. Eq. 10 states that the between-school component of Y is a linear function of a weighted between-school component of X and an error term, v_j. In this case the error term, v_j, is correlated with $\bar{X}_{.j}$. It can be defined as follows (rewriting eq. 10):

$$v_j = \bar{Y}_{.j} - b_{yx} (\bar{X}_{.j} - \bar{X}_{..}) \tag{11}$$

where $\bar{X}_{..}$ is the grand mean on X. The reader familiar with the ANCOVA model will recognize v_j as the "adjusted Y mean." The deviation of the adjusted mean for the jth school from the grand mean of Y is interpreted as the effect of the jth school on Y, free of compositional differences on X. In contrast, the term $b_{yx} (\bar{X}_{.j} - \bar{X}_{..})$ represents the effect of school composition (see Duncan et al., 1972, p. 195; Hauser, 1971, p. 17).

From eq. 10 we conclude that the effect of $\bar{X}_{.j}$ on $\bar{Y}_{.j}$ must operate via one or both of the components of $\bar{Y}_{.j}$. We may therefore write an equation for each of these components in terms of $\bar{X}_{.j}$ as follows (for convenience we set $d_j = b_{yx} (\bar{X}_{.j} - \bar{X}_{..})$):

$$v_j = b_{v\bar{X}} \bar{X}_{.j} + w_j \tag{12.1}$$

$$d_j = b_{d\bar{X}} \bar{X}_{.j} \tag{12.2}$$

Solving for $b_{v\bar{X}}$ and $b_{d\bar{X}}$ we have

$$\begin{aligned} b_{v\bar{X}} &= \mathrm{Cov}(v_j, \bar{X}_{.j}) / \mathrm{Var}(\bar{X}_{.j}) \\ &= \mathrm{Cov}(\bar{Y}_{.j}, \bar{X}_{.j}) / \mathrm{Var}(\bar{X}_{.j}) \\ &\quad - b_{yx} \underline{\mathrm{Var}(\bar{X}_{.j})} / \mathrm{Var}(\bar{X}_{.j}) \\ &= b_{\bar{Y}\bar{X}} - b_{yx} \end{aligned} \tag{13}$$

and

$$\begin{aligned} b_{d\bar{X}} &= \mathrm{Cov}(d_j, \bar{X}_{.j}) \,/\, \mathrm{Var}(\bar{X}_{.j}) \\ &= b_{yx}\, \mathrm{Cov}(\bar{X}_{.j}, \bar{X}_{.j}) \,/\, \mathrm{Var}(\bar{X}_{.j}) \\ &= b_{yx}\, \mathrm{Var}(\bar{X}_{.j}) \,/\, \mathrm{Var}(\bar{X}_{.j}) \\ &= b_{yx} \qquad (14) \end{aligned}$$

Summing these two indirect effects, we have the identity $b_{\overline{YX}} = b_{yx} + (b_{\overline{YX}} - b_{yx})$. We have pointed out that in order for a method to identify school effects it must focus on the difference between a model which allows for between-school effects and one which does not. The critical difference is identified as the difference between the coefficients $b_{\overline{YX}}$ and b_{yx}. The ANCOVA model defines school effects in a manner consistent with this requirement. The model states that the effect of $\bar{X}_{.j}$ on $\bar{Y}_{.j}$ operates via two components, v_j and $b_{yx}(\bar{X}_{.j} - \bar{X}_{..})$. In effect, this decomposes the between-school effect into two parts, one representing the spurious effects of school composition and one representing the effects of unmeasured noncompositional factors, e.g. contextual effects.

Contextual Variables and School Effects

In general analysts of school effects have made use of measures of group composition to indicate the presence of students in particular types of schools. The use of compositional or "contextual" variables is not uncommon in sociological studies in pursuit of group effects (see Riley, 1964, pp. 1017-1019). The following type of structural equation might be estimated using this method:

$$Y_{ij} = a'_{YX} + b'_{YX} X_{ij} + b'_{Y\bar{X}} \bar{X}_{.j} + e'_{ij} \qquad (15)$$

Eq. 15 states that Y is a linear function of two correlated variables, X_{ij} and $\bar{X}_{.j}$, and an uncorrelated error term. As above, this model assumes the error term has a zero population mean, the variance of the error is constant over levels of the independent variables and the values of the error are mutually unrelated. Here the effect b'_{YX} represents the amount Y would change for a given unit change in X_{ij} independent of $\overline{X}_{.j}$ (or group membership), and the effect $b'_{Y\overline{X}}$ is the amount Y would change for a unit shift in $\overline{X}_{.j}$ independent of X_{ij}.[7]

Eq. 15 may be rewritten, by adding and subtracting the term $\overline{X}_{.j}$ to X_{ij} as follows:

$$Y_{ij} = a'_{YX} + b'_{YX}(X_{ij} - \overline{X}_{.j}) + b'_{YX}\overline{X}_{.j} + b'_{Y\overline{X}}\overline{X}_{.j} + e'_{ij} \qquad (16)$$

Grouping terms in eq. 16 produces

$$Y_{ij} = a'_{YX} + b'_{YX}(X_{ij} - \overline{X}_{.j}) + (b'_{YX} + b'_{Y\overline{X}})\overline{X}_{.j} + e'_{ij} \qquad (17)$$

Writing the normal equations for eq. 17 and solving for the parameters produces the following least squares estimates:

$$b'_{YX} = \mathrm{Cov}(Y,x) / \mathrm{Var}(x) \qquad (18)$$

$$b'_{YX} + b'_{Y\overline{X}} = \mathrm{Cov}(Y,\overline{X}) / \mathrm{Var}(\overline{X}) \qquad (19)$$

This means that $b'_{YX} = b_{Yx} = b_{yx}$ and that $(b'_{YX} + b'_{Y\overline{X}}) = b_{Y\overline{X}} = b_{\overline{YX}}$. It follows therefore that the coefficient for the contextual variable $\overline{X}_{.j}$ in eq. 15 is equal to $b_{\overline{YX}} - b_{yx}$. This points out an interesting correspondence between the ANCOVA model and the use of compositional variables. First, both include a component which equals the within-school effect, b_{yx}. Similarly, both models decompose the between-school effect into two correlated components, as follows:

$$b_{\overline{YX}} = b_{yx} + (b_{\overline{YX}} - b_{yx}) \qquad (20)$$

$$= b'_{YX} + b'_{Y\overline{X}} \qquad (21)$$

The school context effects defined by the two methods are equivalent.[8]

Although we have for convenience considered only the simplest case in which X affects Y, our results generalize to the multivariate case, i.e. multiple X's.

We need emphasize the fact that while these identities hold, the use of a set of dummy variables to represent schools in ANCOVA rather than the X school means will ordinarily allow schools to explain more variance in Y net of individual-level variables. This can be seen by observing the fact that there is a residual term in the expression for v_j in eq. 12.1. To the extent that there is variation in this residual, the variance explained by net school effects using a set of dummy variables cannot entirely be accounted for by contextual effects alone. The use of dummy variables then allows one to obtain an upper limit on the amount of variance potentially attributable to school differences. It does not assist in specifying the sources of school-to-school variation in addition to compositional factors which account for the net variance attributable to schools in Y. We should note, however, that the use of dummy variables becomes intractable when the number of schools is large.

Standardizing School Effects

Researchers often prefer to express the effects of school context in a standard metric, especially when comparisons are made with the effects of other variables within the same equation. The principle of standardization discussed by Blalock (1968, p. 190) can be applied in straightforward fashion in the present case, since it simply involves the multiplication of each of the relevant regression coefficients by a ratio of standard deviations. If we standardize the coefficient $b_{\overline{YX}}$ in eq. 20

we have

$$\sigma_{\overline{X}} / \sigma_{\overline{Y}} \;\; b_{\overline{XY}} = \sigma_{\overline{X}} / \sigma_{\overline{Y}} \;\; [\, b_{yx} + (b_{\overline{YX}} - b_{yx})] \qquad (22)$$

The term on the right-hand side of eq. 22, $\sigma_{\overline{X}} / \sigma_{\overline{Y}} (b_{\overline{YX}} - b_{yx})$, represents the contextual component of the effect of $\overline{X}_{.j}$ on $\overline{Y}_{.j}$. In order to produce the standardized contextual effect it is necessary to multiply this effect by the effect of $\overline{Y}_{.j}$ on Y_{ij}, and since the latter effect is unity, the standardized effect is $\sigma_{\overline{Y}} / \sigma_{Y} \; b_{Y\overline{Y}} = \sigma_{\overline{Y}} / \sigma_{Y}$. When we multiply the contextual component of $b_{\overline{YX}}$ by this ratio we have

$$\sigma_{\overline{Y}} \, \sigma_{\overline{X}} \; / \; \sigma_{Y} \, \sigma_{\overline{Y}} \quad (b_{\overline{YX}} - b_{yx}) \qquad (23)$$

$$= \quad \sigma_{\overline{X}} / \sigma_{Y} \;\; (b_{\overline{YX}} - b_{yx}) \qquad (24)$$

The expression in eq. 24 is the standardized contextual effect. This coefficient is a path coefficient and can be interpreted using procedures for the decomposition of effects (see Alwin and Hauser, 1975).

Some Complications

Up to this point we have assumed 1) perfect measurement of all variables in the school effects model, 2) full knowledge of the schooling process within schools and proper assumptions about the nature of unknown sources of variation, and 3) no explicit assignment of students to schools on the basis of the dependent variables involved, e.g. achievement scores. These assumptions may be unrealistic in a variety of situations, and in extreme cases may prevent any inference about school effects. We briefly consider these problems here under the headings of measurement error and selection bias with reference to the applicability of the above model to the study of school effects.

Measurement error--Random measurement error in the covariates (the X's) of the ANCOVA model may produce an unknown degree of bias in the estimated effects of X (or a set of X's) on Y, and some correction for unreliability may be desirable (Lord and Novick, 1968). Taking the simple model developed in eq. 2, random measurement error may affect either or both of the components of X. If the school means on X are infalliable (measured without error) our estimate of the between-school effect of X on Y in eq. 8 will not be biased. In this case we would be concerned only with bias in the within-school slope. It can be seen that random error variance in our estimate of the variance of $X_{ij} - \overline{X}_{.j}$ produces a downward bias in the estimate of b_{yx} in eq. 7. This means that random measurement error in this simple case produces an over-estimate of the effects of schools. In reality we have no way of knowing whether the school means are falliable. In the case where measurement error affects both within and between components, the school effects may be underestimated. Werts and Linn (1971a) show that generally when the within-school coefficient in eq. 2 exceeds the between-school coefficient (by at least a factor of 2) or they are of opposite signs, the estimate of the school effect will be conservative, and reliability corrections will enhance the school effects. In the case of multiple fallible covariates the effects of unreliability are not well-understood, but there appear to be few substitutes for good measurement of theoretically relevant variables.

Selection bias--Our analysis has not taken the within-school aspect of the model as problematic. We have assumed knowledge of the variables affecting Y within schools, that unknown sources of variation in Y are uncorrelated with known sources (in order to produce least squares estimates of the parameters), and that the conditional variance of Y given X (or some set of X's) is independent of levels of X. Taking our simple model in eq. 2 again, we have assumed that schools differ only in their relative levels of X, not in the amount of variability of X within schools. This assumption violates what we know to be true in some classes of school phenomena. For example, colleges explicitly select "qualifiers" from an applicant pool in many cases. This has the effect of reducing the standard deviation within the selected group (Lord and Novick, 1968, pp. 141-142). This may be true to a lesser extent in high schools, where levels of achievement or college-going among students may attract families to a residential area. In U. S. public schools "tracking" among schools is a relatively rare phenomenon. Still, selection may be incidental (Lord and Novick, 1968, p. 142) if not explicit, and selection bias may result in the estimates of the parameters in eq. 2 (see Hannan and Burstein, 1974, pp. 382-384). The problem for the model developed above is twofold. Selection may result in heterogeneity of within-school effects or it may result in the inappropriateness of the assumption of uncorrelated disturbances. As a consequence, differences in within- and between-school coefficients may be observed where none in fact exist (Hauser, 1974), or differences which actually occur may be hidden. The problem may produce inferential errors of an unknown degree, but we imagine that in general it is less of a problem where considerable variability occurs within schools in factors producing outcomes of interest, and when such effects are large. Finally, in some situations the problem of measurement error and selection may combine to produce a welcome result. Goldberger (1972) has shown that under certain types of explicit selection, where cases are assigned to groups on the basis of a fallible variable, unbiased estimates of group effects can be obtained using the ANCOVA model without unreliability corrections. It is unclear whether Goldberger's assignment rule is applicable to the phenomena confronted by investigators of school effects, but it suggests that measurement error and selection problems are not intractable.

Negative Contextual Effects

An aspect of the problem of assessing school context effects which has caused renewed interest in the subject is stimulated in part by James Davis' (1966) formulation of the "frogpond" effect, originally used in explaining the aspirations of college students. A frogpond effect can be viewed as a type of contextual effect which, by virtue of its effects on other variables, has an overall negative impact on the outcomes of schooling. Such effects have been called "frogpond" effects because they imply, for example, that high ability students in less competitive environments, viz. schools with lower average ability, have advantages over students of similar ability in more highly competitive environments. The metaphor that "it is better to be a big frog in a small pond than a small frog in a big pond" is seen to apply (Davis, 1966; Meyer, 1970). Consider the following structural equation representation of the problem:

$$Y = a + b_{YX_1} X_1 + b_{YX_2} X_2 + b_{Y\overline{X}_1} \overline{X}_1 - b_{Y\overline{X}_2} \overline{X}_2 + e \quad (25)$$

where

Y = some outcome of schooling, e.g. college plans

X_1 = student SES

X_2 = student ability

$\overline{X}_1$ = average student SES of school

$\overline{X}_2$ = average student ability of school

Note that while the effect of school SES context on Y is positive, the effect of school ability context is negative. According to frogpond theorists (e.g. Davis, 1966), this happens because schools higher in average SES and average ability are also higher in the existing competitive atmosphere, which depresses individual student performance and aspirations net of individual ability and SES.

Since the concept of competitiveness has been invoked primarily as an interpretation of the effects of variables like $\bar{X}_1$ and $\bar{X}_2$, we may conceptualize it as follows: $C = d_1 \bar{X}_1 + d_2 \bar{X}_2$. (We have omitted any other determinants of the concept in order to simplify the discussion somewhat.) Naturally, both contextual factors, $\bar{X}_1$ and $\bar{X}_2$, may influence the outcome independently of the competitiveness construct, as in Figure 1.

Insert Figure 1 about here

When the competitiveness variable is added to the model in eq. 25, the following structural equation results:

$$Y = a + b_{YX_1} X_1 + b_{YX_2} X_2 + b^*_{Y\bar{X}_1} \bar{X}_1 + b^*_{Y\bar{X}_2} \bar{X}_2 - b_{YC} C + e \quad (26)$$

Note that the within-school effects of X_1 and X_2 are not at issue and are not affected by the inclusion of competitiveness in the model. Since C is a function of $\bar{X}_1$ and $\bar{X}_2$ (see above), eq. 26 may be rewritten as

$$Y = a + b_{YX_1} X_1 + b_{YX_2} X_2 + (b^*_{Y\bar{X}_1} - b_{YC} d_1) \bar{X}_1 + (b^*_{Y\bar{X}_2} - b_{YC} d_2) \bar{X}_2 + e \quad (27)$$

Eq. 27 indicates that if eq. 25 is correct, it must be the case that $b^*_{Y\bar{X}_1} > b_{YC} d_1$ and $b^*_{Y\bar{X}_2} < b_{YC} d_2$.

Clearly, if the competitiveness of the school context operates in the manner described above, the omission of $\bar{X}_2$ from eq. 25 will suppress the effect of $\overline{X}_1$ there. This is consistent with Meyer's (1970) finding that when the positive net effect of school socioeconomic status is considered in light of the negative effects of school context on college-going plans, the school status effect is enhanced. His conclusion is that negative indirect effects of school context on college plans, i.e. frog-pond effects, contribute to the suppression of significant positive direct effects of school socioeconomic status (Meyer, 1970, pp. 68-69). This formulation is also consistent with Nelson's (1972) finding that net of high school grade performance (a proxy in his analysis for competitiveness), the socioeconomic status of the school is more highly related to students' college plans. This leads him to conclude that the role of school context has been masked by the omission of grade performance in previous studies (Nelson, 1972, p. 147). Judging from the conclusions of both Meyer and Nelson, it is essential that studies of school effects adequately specify both the relevant sources of between-school variation and the within-school factors which serve to transmit the effects of school-level variables. Unfortunately, the competitiveness concept has never been directly measured in school effect studies, and only indirect evidence has been brought to bear on the frogpond hypothesis. Short of direct measurement of competitiveness, the pattern of effects described in eq. 25 is taken to be the reduced-form of the frogpond model implicit in the discussions of frogpond theorists. Studies using this basic model (e.g. Alexander and Eckland, 1975; Jencks and Brown, 1975; Hauser et al., 1976) have found only weak support for the frogpond hypothesis.

Illustration

Using data on high school seniors representative of youth in schools in the state of Washington, 1964-66, we address two major questions raised by the literature on school effects. First, we address the question of the magnitudes of net school context effects on educational and occupational aspirations. Just as we recognize the inevitability of different points of view on the interpretation of data on school effects, we see the need for the replication of past analyses so that both the findings and conclusions of earlier research may be subjected to the rigorous requirements of external validity. Our first objective then is to assess the magnitudes of school effects. Second, we examine the possibility that negative contextual (or frogpond) effects might be implicated in these data in suppressing the positive effects of other contextual factors. The model guiding the analysis includes measures of socioeconomic background, academic ability, grade performance, curriculum placement, social influence factors, and school means on all exogenous variables (sex, ability and SES variables). We begin with a discussion of our theoretical model followed by a description of the source of the data and our findings.

Our analysis replicates a school process model set forth by Hauser et al. (1976) which incorporates school context as a source of variation in a number of school process variables, including academic certification (grades and curriculum placement), social influence (teachers' and parents' encouragement to attend college and friends' college plans), and aspirations. The model is presented in diagram form in Figure 2. This model states that academic certification--grades (G) and curriculum placement (C)--depends upon individual differences in academic ability (Q), sex (S), socioeconomic background--father's (V) and mother's education (M), father's occupational status (X) and parental income (I)--and school context.

Insert Figure 2 about here

Social influence factors--teachers' (T) and parents' encouragement (P) to attend college and friends' college plans (F)--depend upon academic certification, academic ability, sex, socioeconomic background and school context. Finally, educational (E) and occupational aspirations (J) depend upon all the previous variables. Since all equations estimated in this model include school context, the remaining parameters may be considered within-school parameters and estimates of processes which occur within schools (see Hauser, 1971; Heyns, 1974). Our concern with the present model focuses attention solely on the effects of contextual factors. A discussion of the within school process is presented elsewhere (Alwin, Otto and Call, 1976).

If the between-school regression coefficient and the within-school coefficient are the same, i.e. $b_{\overline{YX}} = b_{yx}$, there can be no school context effect. The issue of school effects can therefore be resolved quite easily (see Schuessler, 1971 pp. 210-213). In the present analysis we estimate several models of the form:

$$Y = a + \sum_k b_{YX_k} X_k + \sum_k b_{YX_{kj}} \overline{X}_{kj} + \sum_p b_{YX_p} X_p + e \qquad (28)$$

where, as above, j varies over schools, k varies over all exogenous individual-level variables and p varies over all endogenous individual-level independent variables. For each dependent variable in the model we estimate a series of equations, beginning with the reduced-form of the model and adding intervening variables in sequence (see Alwin and Hauser, 1975). It is often the case that the $\overline{X}_k$'s are highly correlated in any given body of data, and it is impractical to make the simple comparisons of within- and between-school regression coefficients. This problem may be solved by examining the increment to explained variance introduced by the addition of the $\overline{X}_k$'s to an equation containing the X_k's and the X_p's.[9]

The analysis uses data from three separate samples representative of youth in schools in the state of Washington, 1964-66.[10] Stratified samples of schools were obtained covering all geographical areas of the state of Washington. Classroom-administered questionnaires obtained from students attending sampled schools were one of several sources of data. Additional information from subsamples of students was gathered through personal interviews with students and from parents using mail questionnaires. Information was obtained from school records with the assistance of school counselors (see Slocum, 1966). Our analysis uses data from the seniors in the three samples (N = 4,303).[11] Most of the variables used here are derived from the school questionnaire. Father's (V) and mother's education (M) (coded in years of schooling completed) and father's occupational status (X) (coded in the metric of Duncan's SEI) were obtained from student reports on the questionnaire. Parental income (I) comes from parental sources (sample 1) and student interviews (samples 2 and 3) and is coded in $100 units for the midpoints of the response intervals used. Two school-related variables, grades (G) and academic ability (Q), were obtained from school records for students selected for intensive study in all samples. Grades are expressed in grade point averages (4 point scale) and academic ability in the metric of I. Q. scores. The curriculum placement (C) of the student was constructed from student reports on the questionnaire of courses taken during high school. This variable is given a high score if the student had completed a course of study enabling him or her to enter the University of Washington. Three measures of social influence--perceived teachers' (T) and parents' (P) encouragement to attend college and friends' college plans (F)-- were obtained from the school questionnaire and are expressed as binary variables. Finally, the student's college plans (E), expressing whether he or she desired attendance at a four-year college or university, and occupational status aspirations (J), coded in the metric of Duncan's SEI, were both obtained from the school questionnaire. The variables are for the most part comparable to those used by Hauser et al. (1976) in a similar study of 1957 Wisconsin high school seniors. The Wisconsin data provide a useful bench-mark against which to evaluate the integrity of the Washington data. Table 1 presents the descriptive univariate statistics for the two populations.

Insert Table 1 about here

The Washington data contain students from families with higher levels of parental education, income and occupational status. Washington students have higher levels of educational and occupational aspirations; they perceive more college-going intentions among their friends, less encouragement from their parents and about the same level of encouragement from teachers; and they gain placement in college preparatory curricula in the same proportions as do Wisconsin students nearly a decade earlier. The average ability scores in the two populations are roughly comparable, but we connot compare the grades of the two populations owing to metric differences. Some of the differences described here can be attributed to the fact that different age cohorts are involved. For example, younger cohorts can be expected to have better educated parents (Hauser and Featherman, 1976), parental incomes are subject to inflation, and college going was more common in the mid-sixties than in the mid-fifties. However, we suspect that some of the differences may be due to differing definitions of the populations, data coverage and the particular measures of variables used.

Table 1 also indicates that most of the variance in variables in both populations lies within schools. This mitigates any expectation of the importance of school context effects relative to processes which operate within schools. Socioeconomic background variables typically have the largest proportions of between-school variance. The between-school standard deviations are roughly comparable in the two populations. With the exception of grades, in the remaining variables the proportions of between-school variance are less in Washington than in Wisconsin. The Wisconsin measure of grade performance, rank in high school class, inherently reduces the between-school variation in that variable, while there is no such control over between-school differences in the allocation of grades in the Washington data. It is clear from Table 1 that in general the removal of the between-school variance from the total variances results in much closer comparability in the within-school distributions of variables in the two populations. These results generally support our view that the Washington data come from a population comparable in many respects to the population studied by Sewell and his associates (Hauser et al., 1976).

Additive School Effects in the Washington Data

The following discussion focuses solely on the question of additive school effects in the data described above. We have not examined the possibility of interaction effects by school. Hauser et al. (1976) undertake an exhaustive search for interactions in the Milwaukee data originally analyzed by Sewell and Armer (1966) and find none. We use this as the basis for our assumption of the within-school homogeneity of slopes, i.e., additivity. The within- and between-school correlation matrices are given in Table 2.[12] The correlations among the specific compositional measures in the matrix, e.g., measures of school socioeconomic context, require a cautious approach to the analysis of school effects. As we indicated above (see fn. 9) our inclusion of compositional measures was dictated by standards for inclusion which guard against problems of multi-collinearity.

Insert Tables 2 & 3 about here

Table 3 presents the gross and net contributions of school contextual variables to the variance in the school process variables. The first line presents the proportions of variance occuring between schools in the sample. Even-numbered lines thereafter present the proportions of variance explained by the individual-level school process variables, and odd-numbered lines thereafter contain the proportions of variance explained in an outcome by adding the school context variables. The net contributions of the school context variables are given in parentheses. These quantities are formed by subtracting the coefficient in an even-numbered line from the coefficient in the odd-numbered line immediately below. For example, 11.83 percent of the sample variance in grade point average (G) lies among schools (see line 1 of Table 3). After these differences are adjusted for prior differences in socioeconomic background, academic ability and sex the percentage of variance in grades attributable to school context is 1.37. We conclude from this that most of the differences in the allocation of grades among schools are due to differences in socioeconomic background, academic ability and sex and the importance of these variables for grades within schools. Similarly, most of the between-school variance in curriculum placement is accounted for by prior differences in SES factors, ability and sex. About 3.4 percent of the total variance in curriculum placement (C) is accounted for by school context factors. Obviously, school context figures more importantly for individual differences in curriculum placement than grades, accounting for about one-tenth of the explained variance in curriculum placement and less than one-twenty-fifth of the explained variance in grades.

The net contribution of school context to variation in the social influence variables is most pronounced for friends' college plans (F), contributing about 3 percent to the variance in F or about one-seventh of the explained variance. This represents a substantial reduction in the gross between-school variance. Most of this reduction is due to controlling differences on socioeconomic background, ability and sex, and very little to controlling grades and curriculum. The contributions of school context factors to variance in teachers' and parents' encouragement are negligible when SES variables, ability and sex are controlled, and essentially unaffected when grades and curriculum are added to the model. There is very slight support for a "frogpond" effect on parental encouragement. The contribution of school context factors increases when grades and curriculum placement are added to the model (compare lines 3 and 5), suggesting that there is a negative indirect contextual effect via grades and curriculum which is not apparent when these intervening variables are not in the equation. This effect is trivially small, accounting for less than one percent of the variance in parental encouragement.

Most of the early literature on school context focused on college plans as the dependent variable. In some ways it is unfortunate that the debate was initiated over this type of measure. First, it would appear that achievement criteria, e.g., academic achievement and educational attainment, are more valid indicators for comparing the effectiveness of schools. Second, the between-school variance in college plans tends to be small relative to other possible outcomes. Our own analysis, however, is limited ultimately to college plans and occupational status aspirations. The between-school component of college plans accounts for 7.62 percent of the variance in the variable in these data. The corresponding figure for occupational status aspirations is 3.91. Controlling for differences in socioeconomic background, academic ability and sex and for the fact that these variables affect aspirations within schools, reduces the contribution of school context to 1.63 and 1.69 percent of the variance in college plans and occupational status aspirations, respectively. Taking intervening school and social influences into account as well reduces the net contributions in these variables to 1.15 and 1.12 percent of the variances involved (see line 7 of Table 3), about one-fortieth of the explained variance in college plans and about one-twenty-fifth of the explained variance in aspirations.

Summarizing the results up to this point, school context accounts for scarcely more than 1 percent of the variance in five of the seven outcomes in the model. Exceptions are the curriculum placement of students into college vs. non-college preparatory tracks and the perceived college plans of friends, which depend to a slightly larger extent on school context. There is no evidence here for strong effects of school context factors on outcomes of schooling. We now turn to an examination of the contextual effects themselves in order to determine whether "frogpond" phenomena contribute to these small effects.

Frogpond Effects

Contextual effects which are negative have come to be referred to as "frogpond" effects because they imply that high ability students in less competitive environments (e.g. schools with lower mean ability) have advantages over students of equal ability in more highly competitive environments. The metaphor, that "it is better to be a big frog in a small pond than a small frog in a big pond," is seen to apply (Davis, 1966; Meyer, 1970; Nelson, 1972). We now investigate the applicability of the metaphor to the contextual effects found in the Washington data. The regression equations which produced the coefficients of determination in the odd-numbered lines of Table 3 are presented in standard form in Table 4. Both reduced-form and structural coefficients are given in the table for models of the form discussed above. The equations contain both individual-level variables and school means for the exogenous variables.[13] As we have noted earlier, we have been cautious in our inclusion of the contextual variables because they are highly correlated (see fn. 9). We will confine our attention here primarily to the contextual effects, referring to the within-school effects for comparison purposes only. A full interpretation of the within-school effects is presented elsewhere (Alwin, Otto and Call, 1976).

The first two equations in Table 4 present estimates for the regression of grades (G) and curriculum placement (C) on sex (S), socioeconomic background (B), academic ability (Q), and the school context variables. Following Heise (1972) we have blocked the coefficients for V, M, X, and I, since the inclusion of even moderately redundant sets of variables provides a misleading picture of relative effects (Gordon, 1968). Note also that our criteria for the inclusion of the compositional measures result in a rather refined set of contextual variables in each equation.[14] The equation in line 1 of the table indicates that two contextual variables,

----

Insert Table 4 about here

----

the sex composition of the school and the ability composition, affect the student's grade point average. The proportion of males in the school ($\overline{S}$) decreases grade point average, and the average ability ($\overline{Q}$) of students in a school also has a depressing effect on grades. The latter effect is clearly consistent with the frog-pond interpretation of school effects reported by Meyer (1970) and Nelson (1972). In both instances the context effect is less in relative magnitude than the corresponding individual-level variable's effect. The sex and ability composition of the school also affect curriculum placement, as does the socioeconomic context (as average father's occupational status). Again, schools with students of higher average ability tend to depress the likelihood that students will be placed in college preparatory programs. This is again consistent with the notion advanced by Meyer (1970) that the competitiveness of the school depresses individual performance on academic tasks, in addition to having positive effects on other variables. The latter type of effect of school context is illustrated by the substantial effect of socioeconomic context ($\bar{X}$) on curriculum placement. This effect accounts in large part for the result discussed with reference to curriculum placement in Table 3. Finally, the effect of the sex composition of the high school class is judged to be nonsignificant.[15]

There is a small effect of socioeconomic context on teachers' encouragement (T) tapped by average father's education. The increase in this small effect when the academic certification variables (G and C) are added to the equation (line 4) indicates the suppressing effect of school context via grade point average and curriculum placement, but the effect is trivially small, accounting for very little variance in teacher's encouragement (compare lines 3 and 5 in Table 3). It is not surprising to find that the likelihood that one's friends plan to attend college is decreased in higher average ability schools net of socioeconomic context and individual-level variables (S, B, and Q). Nor is it unusual that the likelihood of having college-bound friends is dependent upon the socioeconomic context of the school (see Hauser et al., 1976). The latter effect is stronger than the combined effects of socioeconomic background variables and the effects of academic ability, grades and curriculum. There is also a small effect of the sex composition of the school on friends' plans, favoring male dominated classes. At the same time, the higher the ability context of the school, the more likely are parents to encourage their children to attend college (or so it is perceived by the students), and the higher the socioeconomic context of the school, the less likely are parents to offer encouragement. The effects of school context on parental encouragement are enhanced slightly when grades and curriculum are controlled (line 6), indicating the presence of small suppressor effects. It is clear that the encouragement of parents and the reported plans of friends are affected in opposite ways by the school context. These findings suggest that parents in higher (socioeconomic) status schools play a secondary role relative to the stronger influence of peers, whereas in higher ability schools peers are a less important source of social influence than parents. As we indicated earlier, the encouragement of teachers to attend college does not depend on either the socioeconomic or ability context of the school to any nontrivial extent. This is consistent with other findings which indicate the relatively meritocratic character of teachers' encouragement (see Hauser et al., 1976).

The equations in lines 9-11 of Table 4 present our results for college plans (E). These findings provide no support for an effect of the socioeconomic composition of the school, as others have observed (Rogoff, 1961; Michael, 1961; Turner, 1964; Wilson, 1959; Boyle, 1966; Meyer, 1970). There is a negative effect of the ability context on college plans (line 9). Contrary to an expectation based on Nelson's (1972) claim, very little of the effect of ability context (-.134 + .107 = .027) is mediated by grades and curriculum. The ability context effect clearly does not suppress any effect of the socioeconomic context of the school. The only other effect of school context on college plans is an insignificant effect of sex composition. Finally, lines 12-14 present the coefficients for occupational status aspirations. There is a positive effect of the income composition ($\overline{I}$) of the school and a negative effect of the ability context ($\overline{Q}$). Neither effect is substantially altered by the inclusion of the academic (line 13) and social influence (line 14) variables.

Summary and Conclusions

The purpose of this empirical analysis has been to assess the relative magnitudes of school context effects on several school process variables in a sample of high school seniors, and to examine the extent to which negative contextual or frogpond effects occur in the data. We have analyzed the combined data from three separate samples which are geographically representative of students in schools in the state of Washington, 1964-66. Our approach to the data has been guided by a theoretical model set forth by Hauser et al. (1976) in which school context is explicitly represented as an exogenous source of variation in academic certification (grades and curriculum), social influences (the influence of teachers, parents and friends), and aspirations (college plans and occupational status aspirations). These exogenous effects are considered net of sex, ability and socioeconomic background factors. Our model also examines the intervening process by which these effects are produced inasmuch as the model set forth accounts for it. Our analysis is at slight variance with the approach of Hauser et al. (1976), but consistent with other analyses (e.g. Alexander and Eckland, 1975), by including only measures of school composition on the socioeconomic background variables, academic ability and sex.

We have studied only the additive effects of schools in this body of data and have assumed additivity on the basis of the results of the extensive search for interactions carried out by Hauser et al. (1976) in the Milwaukee subsample of the Wisconsin data. After a preliminary analysis demonstrating the comparability of the Washington data to the Wisconsin sample of 1957 high school seniors, we presented the gross and net contributions of school context factors to variation in individual-level outcomes. Then we examined the patterns of contextual effects more closely in an effort to detect the presence of "frogpond" effects. We found no evidence for strong contextual effects on the school process variables, although such effects are clearly present. School contexts account for scarcely more than one percent of the variance in five of the seven outcomes studied. They account for about three percent of the variance in curriculum placement and friends' college plans. Relative to the ability of within-school processes to account for variation in these variables, school context is most important in determining friends' college plans, followed in order of importance by curriculum placement, parental encouragement, grades, occupational aspirations, college plans and teachers' encouragement. This was further borne out by our examination of the contextual coefficients in Table 4. The effect of school socioeconomic context on friends' college plans is stronger than the combined effects of socioeconomic background, and the effects of ability, grades and curriculum. The effect of socioeconomic context on curriculum placement is larger than the effects of sex and background, but not academic ability. The remaining contextual effects identified in the analysis are quite small as indicated by their small net contributions to explained variance. Despite their size there is considerable support in these data for negative contextual effects. Ability context has small negative effects on grades, curriculum placement, friends' college plans, college plans, and occupational aspirations, but in no case does this serve a suppressing role via the effects on grades and curriculum. There is a very slight suppressor effect observed via grades and curriculum for parental encouragement, but its magnitude is trivial. In other words, there are small "frogpond" effects in these data, meaning negative contextual effects, but they are almost entirely unmediated by intervening variables in the model.

We began this paper with a discussion of the optimism of many sociologists with regard to the importance of school context factors in the development of academic skills, self-conceptions, aspirations and achievements in students. We pointed out that while "school" effects have consistently been shown to be small relative to "schooling" effects, their role has often been exaggerated. There is some danger in overgeneralizing about the magnitudes of "school effects" since, as our analysis shows, this depends upon which outcome variable is being discussed. But with few exceptions, the effects of school differences have been found to be small relative to our ability to explain variation in these variables using within-school models. Any interpretation of school effects must consider this point. Taking college plans, the dependent variable most often treated in analyses of school effects, we find that multiple school context factors account for about one-fortieth of the explained variance in this variable here. This figure ranges in our analysis from one-seventh for friends' college plans to one-eightieth for the explained variance in teachers' encouragement. It is our conclusion that while school context clearly affects the processes within schools, there is little evidence here for meaningful effects on the ultimate outcomes of schools, e.g.

aspirations.

For virtually all the dependent variables considered in this analysis, we can account for from 70 to 95 percent of the between-school variance using the within-school structural equation model. Owing to the fact that little (from 2 to 12 percent) of the total variance in these outcomes lies between schools in any case, we can see no reason to exaggerate the potential importance of school context factors. This suggests that the study of school processes be turned to other matters (see Hauser et al., 1976).

Concluding Comments

The above discussion speaks directly to the issue of between-school specification of school effects models. Insofar as individual-level outcomes are concerned, if either the between-school or within-school part of the model is mis-specified, e.g., the omission of relevant variables, bias may result in both the between- and within-school parameters. This would obviously produce bias in estimated school effects.

There is one type of between-school specification problem which the present discussion has not pursued. This involves the issue of "global" or "structural" characteristics of schools (Coleman et al., 1966). It is possible to indicate the place of school characteristics within the present formulation. Such factors would obviously affect Y via $\overline{Y}_{.j}$, so given the specification for $\overline{Y}_{.j}$ in eq. 10, school characteristics either affect the dependent variable via one of these components or interpret the effects of one of these components. As we have defined them, the compositional and non-compositional factors in the model account for the total variation in the Y school means, but the present model does not in principle exclude such "structural" characteristics. Indeed, it may be useful to specify the causal relationships among structural characteristics of the school and contextual variables so often encountered in studies of school effects (e.g. Alexander and Eckland, 1975). Some studies (e.g. Mayeske et al., 1973; Bidwell and Kasarda, 1975) go no further than accounting for variation in $\overline{Y}_{.j}$. Unless such analyses extend their models to individual-level outcomes their value for studying school effects (as defined here) is questionable. Finally, this presentation has presumed no particular type of research design. While the model discussed here is applicable to cross-sectional data, longitudinal designs may be preferable (Gilbert and Mosteller, 1972; Luecke and McGinn, 1975).

FOOTNOTES

1. The exceptions are Sewell and Armer (1966), Hauser (1969, 1971) and Hauser et al. (1976).

2. A second problem Hauser (1970a) sees with traditional forms of contextual analysis is that it is fallacious to arbitrarily identify residual group effects with social or psychological mechanisms which may be correlated with levels of one of the individual-level variables in the analysis (p. 659).

3. Although this analysis focuses primarily on the effects of high schools, the discussion applies equally to other causally relevant bases of aggregation, e.g. colleges (Alwin, 1976), large cities (Mueller, 1974), or groups in general. Our discussion is also relevant to the pursuit of "structural" effects (Blau, 1960; Tannenbaum and Bachman, 1964), "contextual" effects (Riley, 1964; Lazarsfeld and Menzel, 1969), and effects of "group composition" (Davis et al., 1961).

4. In order to obtain a linear unbiased estimate of b_{YX} we need only assume perfect measurement of X, but if we want to estimate the standardized effect (r_{YX} in this case) as well, an unbiased estimate of Var(Y) is needed and this precludes any measurement error in Y (Lord and Novick, 1968).

5. It is possible to express the effect of X on Y in eq. 1 as a weighted sum of the within- and between-school effects in eq. 2. This is done by weighting the within-school effect by the proportion of the total variance in X which lies within schools and the between-school effect by the proportion of the variance in X which lies between schools (see Duncan et al., 1961, pp. 66-67.)

6. The disturbance in this equation represents nonlinearity in the relationship of the X and Y school means, as well as variation in the Y means for schools with the same X mean.

7. The intercept and error term in eq. 15 are identical to those in eq. 2, so we have designated them in the same way.

8. The possible correspondence of the ANCOVA model with contextual analysis is mentioned by Schuessler (1968) and Fennessey (1968), and the identity has been formally stated by Werts and Linn (1971b). The identification of the group effect with the differences in between-group and within-group regression coefficients is becoming more prevalent (e.g. Cronbach and Webb, 1975).

9. In order to avoid problems resulting from extreme multicolinearity, we add our compositional variables to the equation in a stepwise fashion using the criteria that in order to enter the equation a variable's regression coefficient must have a unit F-ratio or greater and must correlate less than .7 with the variables already in the equation. See Nie et al. (1975).

10. These data were collected by Walter L. Slocum and his associates.

11. Sample 1 contains 3,535 cases (1,185 seniors), sample 2 contains 3,660 cases (1,448 seniors), and sample 3 contains 3,603 cases (1,670 seniors). Among students attending school on the date of the administration of the questionnaire the response rate was very high, although 5 percent of the students selected for the personal interview declined to participate and some 30 percent of the parents did not return the mail questionnaire (see Slocum, 1966).

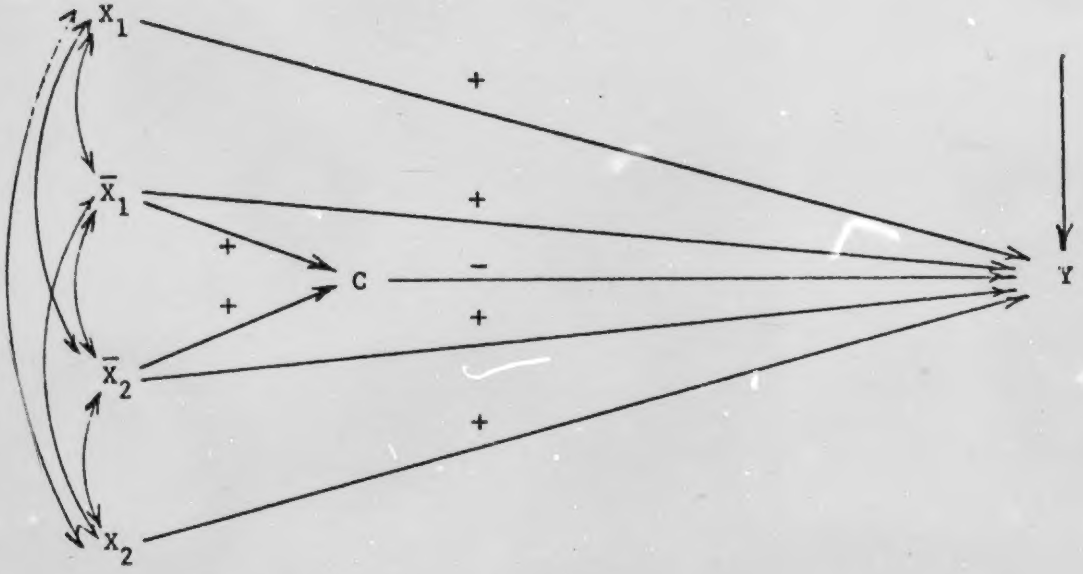
12. The correlation matrix in Table 2 is based on pairwise present data. In three cases the Washington data do not permit the calculation of a correlation because the questions were not asked on the same study. These are r_{TX}, r_{TQ}, and r_{TP}. We obtain these absent correlations from those computed in the Wisconsin data (from personal correspondence with Professor Robert M. Hauser).

13. In table 4 we have presented the exact within-school regression coefficients obtained from a within-school analysis using the within-school correlations in this data. There is a formal equivalence between these coefficients and those obtained using eq. 28.

14. Note that we have included only compositional measures for sex, socioeconomic background and academic ability in our analysis. Estimation of models including compositional measures of grades, curriculum and the social influence variables as well do not substantially alter the findings in Table 4, so we have deleted any presentation of them.

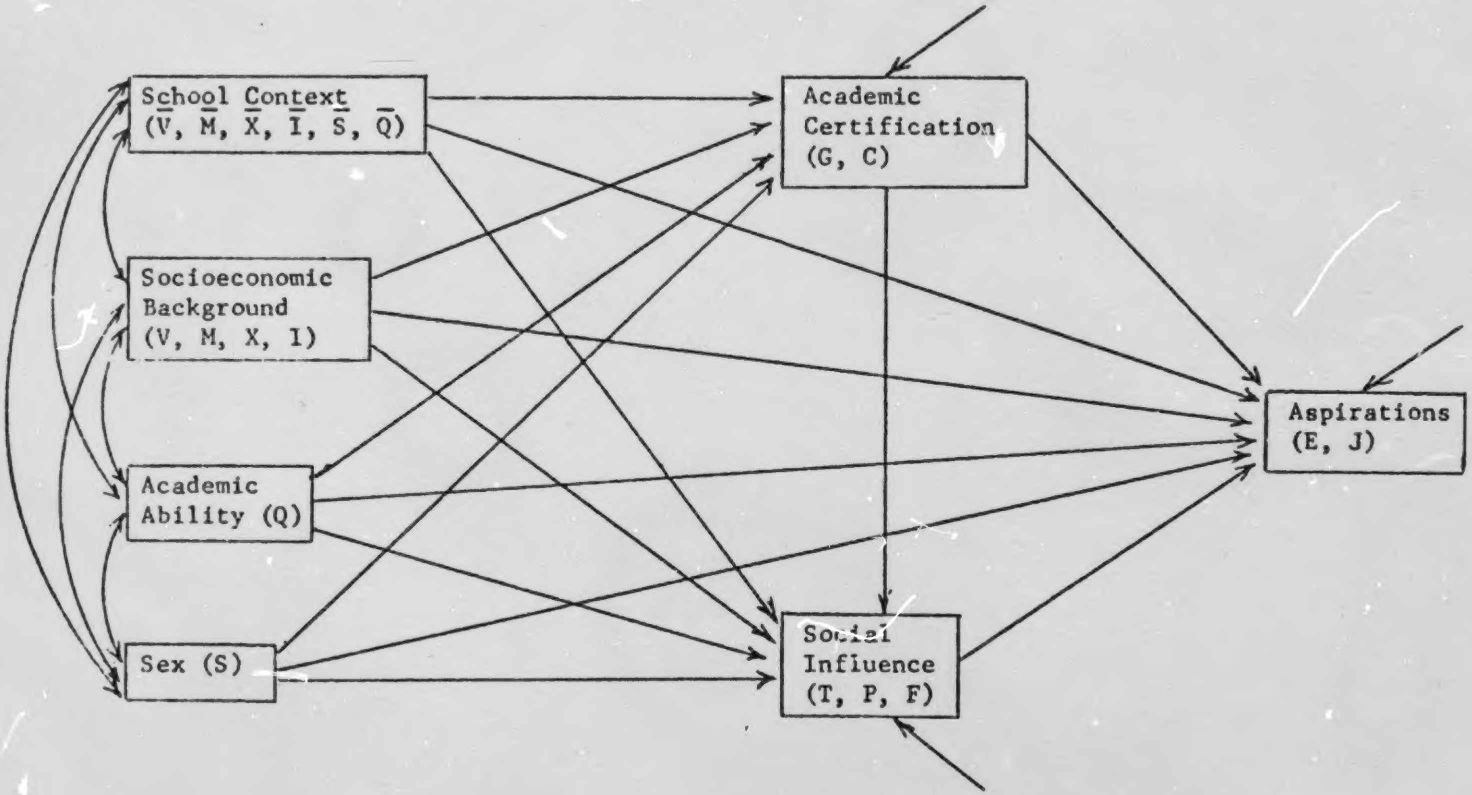
15. Our criterion of statistical significance is really quite arbitrary. As noted above, our sample is a stratified cluster sample and there are unsystematic sources of missing data due to the nature of the research design. Although we do not expect bias in the standard errors for the within-school coefficients by doing so, we have computed standard errors using the formula for a simple random sample. In so doing we may overestimate the standard errors for the contextual effects. We have used N = 1,000 as a conservative estimate of average bivariate sample size in the computation of standard errors.

Figure 1.--Causal diagram for specification of "frogpond" effects.


X_1
$\bar{X}_1$
$\bar{X}_2$
X_2
C
Y
+
+
+
+
−
+
+

Note: See text for description of notation.

Figure 2.--Causal diagram for school process model.


School Context
($\bar{V}$, $\bar{M}$, $\bar{X}$, $\bar{I}$, $\bar{S}$, $\bar{Q}$)
Socioeconomic Background
(V, M, X, I)
Academic Ability (Q)
Sex (S)
Academic Certification
(G, C)
Social Influence
(T, P, F)
Aspirations
(E, J)

Note: See text for variable abbreviations.

Table 1.--Means and standard deviations for all variables in the analysis: high school seniors (1) in the state of Washington, 1964-66, and (2) in the state of Wisconsin, 1957.

Variable	$\bar{X}$		s		s_W		s_B		P_B		N	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
S	.517	.486	.500	.500	.495	.476	.068	.152	.019	.092	4303	7052
V	12.0	10.0	3.03	3.09	2.76	2.82	1.25	1.26	.171	.166	4064	7052
M	12.2	10.4	2.31	2.98	2.16	2.76	.82	1.13	.125	.143	4106	7052
X	46.2	31.0	28.1	22.4	24.9	20.1	12.9	9.82	.210	.193	2694	7052
I	94.7	59.2	40.0	32.9	35.6	28.2	18.2	17.0	.206	.267	701	7052
Q	108.	102.	13.6	14.5	13.0	13.5	4.0	5.16	.086	.127	490	7052
G	2.40	51.6	.700	18.7	.658	18.2	.241	4.01	.118	.046	1019	7052
C	.594	.603	.491	.489	.461	.450	.169	.192	.118	.154	2615	7052
T	.426	.452	.495	.498	.481	.478	.116	.139	.055	.078	1118	7052
P	.362	.540	.481	.498	.354	.470	.070	.166	.021	.111	337	7052
F	.605	.379	.489	.485	.458	.441	.172	.202	.124	.174	2498	7052
E	.603	.353	.489	.478	.470	.449	.135	.165	.076	.118	4303	7052
J	59.2	50.0	22.4	22.6	22.0	21.1	4.4	7.91	.039	.123	3560	7052

Note: Column (1) in each pair of columns contains data from the Washington samples and column (2) contains data from Wisconsin. The source of the Wisconsin data is Hauser et al. (1976), Table 2. $\bar{X}$ is the sample mean, s is the sample standard deviation, s_W is the within-school standard deviation, s_B is the between-school standard deviation, P_B is the proportion of the total variance between schools, and N is the sample size for data present. Variable abbreviations are as follows: S = sex, V = father's education, M = mother's education, X = father's occupational status, I = parental income, Q = academic ability, G = grade point average, C = college preparatory curriculum, T = teachers' encouragement, P = parents' encouragement, F = friends' college plans, E = college plans, J = occupational status aspirations.

Table 2.--Within and between-school correlations for all variables in the analysis (within-school correlations above the main diagonal and between-school correlations below): high school seniors in the state of Washington, 1964-66.

Variable	S	V	M	X	I	Q	G	C	T	P	F	E	J
S	---	.007	.011	-.015	.018	-.064	-.211	-.131	-.048	-.069	.002	.043	.082
V	-.116	---	.428	.505	.310	.249	.229	.209	.096	.114	.120	.248	.166
M	-.166	.922	---	.302	.267	.180	.181	.144	.104	.135	.146	.238	.145
X	-.233	.931	.934	---	.375	.167	.209	.192	.125*	.268	.175	.221	.158
I	-.083	.779	.700	.870	---	.157	.138	.161	-.026	.209	.098	.171	.083
Q	-.032	.370	.462	.358	.413	---	.541	.445	.360*	.130	.188	.456	.363
G	-.378	.204	.335	.362	.055	.110	---	.457	.417	.256	.248	.518	.378
C	-.169	.814	.760	.898	.737	.461	.458	---	.284	.216	.281	.407	.346
T	-.151	.375	.350	.153*	-.113	.302*	.430	-.002	---	.426*	.211	.331	.291
P	-.256	.735	.762	.788	.707	.469	.518	.696	.379*	---	.157	.255	.232
F	-.026	.793	.795	.887	.535	.177	.509	.611	.566	.557	---	.370	.271
E	-.052	.675	.720	.888	.418	.318	.397	.540	.566	.704	.797	---	.441
J	-.052	.768	.758	.900	.594	.251	.125	.670	.038	.648	.640	.609	---

Note: Variable abbreviations are as follows: S = sex, V = father's education, M = mother's education, X = father's occupational status, I = parental income, Q = academic ability, G = grade point average, C = college preparatory curriculum, T = teacher's encouragement, P = parents' encouragement, F = friends' college plans, E = college plans, J = occupational status aspirations.

*correlation obtained from Hauser et al. (1976) for Wisconsin 1957 high school seniors.

Table 3.--Coefficients of determination expressing gross and net contributions of school context to variance in school process variables: high school seniors in the state of Washington, 1964-66.

Predetermined variables	Dependent variable						
	G	C	T	P	F	E	J
1. H	.1183	.1183	.0549	.0215	.1241	.0762	.0391
2. S,V,M,X,I,Q	.3572	.2917	.1498	.1094	.1286	.2691	.1683
3. S,V,M,X,I,Q,H	.3509	.3256	.1509	.1165	.1681	.2854	.1852
	(.0137)	(.0339)	(.0011)	(.0071)	(.0395)	(.0163)	(.0169)
4. S,V,M,X,I,Q,G,C			.2438	.1409	.1856	.4058	.2422
5. S,V,M,X,I,Q,G,C,H			.2470	.1531	.2167	.4157	.2528
			(.0032)	(.0122)	(.0311)	(.0099)	(.0106)
6. S,V,M,X,I,Q,G,C,T,P,F						.4488	.2724
7. S,V,M,X,I,Q,G,C,T,P,F,H						.4603	.2836
						(.0115)	(.0112)

Note: Variable abbreviations are as follows: H = high school context, S = sex, V = father's education, M = mother's education, X = father's occupational status, I = parental income, Q = academic ability, G = grade point average, C = college preparatory curriculum, T = teachers' encouragement, P = parents' encouragement, F = friends' college plans, E = college plans, J = occupational status aspirations.

Table 4.--Coefficients for variables in structural and reduced-form equations (variables in standard form): high school seniors in the state of Washington, 1964-66.

Dependent variable		Predetermined variable												
		S	B**	Q	G	C	T	P	F	$\bar{S}$	$\bar{V}$	$\bar{X}$	$\bar{I}$	$\bar{Q}$
1.	G	-.178*	.141*	.496*						-.081*				-.094*
2.	C	-.106*	.144*	.402*						.029		.237*		-.082*
3.	T	-.021*	.141*	.358*							.038			
4.	T	.041*	.134*	.177*	.297*	.086					.067*			
5.	P	-.063*	.286*	.079*								-.092*		.075*
6.	P	-.018	.252*	-.058	.188*	.109						-.121*		.103*
7.	F	.013	.169*	.156*						.070*		.254*		-.072*
8.	F	.059*	.127*	.006	.140*	.200*				.083*		.224*		-.041
9.	E	.068*	.203*	.410*										-.134*
10.	E	.147*	.135*	.172*	.348*	.160*				.037				-.096*
11.	E	.136*	.104*	.169*	.295*	.110*	.036	.078*	.196*	.026				-.109*
12.	J	.105*	.118*	.345*									.149*	-.115*
13.	J	.166*	.069	.157*	.231*	.182*							.125*	-.084*
14.	J	.158*	.047	.154*	.179*	.141*	.049	.106*	.126*				.123*	-.095*

Variable abbreviations are as follows: S = sex, B = socioeconomic background, Q = academic ability, G = grade point average, C = college preparatory curriculum, T = teachers' encouragement, P = parents' encouragement, F = friends' college plans, E = college plans, J = occupational status aspirations, $\bar{S}$ = proportion males in school, $\bar{V}$ = school average father's education, $\bar{X}$ = school average father's occupational status, $\bar{I}$ = school average parental income, $\bar{Q}$ = school average academic ability.

**The coefficient represents the blocked effect of V, M, X and I.

*The coefficient is greater than twice its standard error.

Alexander, Karl L. and Bruce K. Eckland

1975 "Contextual effects in the high school attainment process," American Sociological Review 40 (June): 402-416.

Alwin, Duane F.

1976 "Socioeconomic background, colleges and post-collegiate achievements." In W.H. Sewell, R. M. Hauser, and D. L. Featherman (eds.) Schooling and Achievement in American Society. New York: Academic Press.

Alwin, Duane F.

1976 "Assessing school effects: some identities." Sociology of Education 49 (October), In press.

Alwin, Duane F. and Robert M. Hauser

1975 "The decomposition of effects in path analysis." American Sociological Review 40 (February): 37-47.

Alwin, Duane F. and Luther B. Otto

1976 "Schools and aspirations: A re-examination of the frogpond hypothesis." Unpublished paper.

Alwin, Duane F., Luther B. Otto, and Vaughn R. A. Call

1976 "The schooling process in the development of aspirations: a replication." Unpublished paper.

Barton, Allen H.

1970 "Allen Barton comments on Hauser's 'Context and consex," American Journal of Sociology 76 (November):514-17.

Bayer, A.E.
1969 "Neighborhood context and college plans." American Journal of Sociology 75(September):239-244.

Bidwell, Charles E and John D. Kasarda
1975 "School district organization and student achievement." American Sociological Review 40 (February):55-70.

Blalock, Hubert M., Jr.

1968 "Theory building and causal inferences." In Hubert M. Blalock, Jr. and Ann B. Blalock (eds.), Methodology in Social Research. New York: McGraw-Hill.

Blau, Peter M.

1960 "Structural effects," American Sociological Review 25 (April):178-93.

Boyle, Richard P.

1966 "The effects of the high school on students' aspirations," American Journal of Sociology 71 (May): 628-39.

Campbell, Ernest Q. and C. Norman Alexander

1965 "Structural effects and interpersonal relationships." American Journal of Sociology 71(November):284-89.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander Mood, Frederick D. Weinfeld, and Robert L. York

1966 Equality of Educational Opportunity. Washington D. C.: U. S. Government Printing Office.

Cronbach, Lee J. and Noreen Webb

1975 "Between-class and within-class effects in a reported aptitude x treatment interaction: reanalysis of a study by G. L. Anderson." Journal of Educational Psychology 67 (no. 6):717-724.

Davis, James A.

1966 "The campus as a frog pond: an application of the theory of relative deprivation to career decisions of college men." American Journal of Sociology 72(July): 17-31.

Davis, James A., Joe L. Spaeth, and Carolyn Huson

1961 "A technique for analyzing the effects of group composition," American Sociological Review 26 (April):215-26.

Duncan, Otis Dudley

1970 "Partials, partitions, and paths." In E. F. Borgatta (ed.), Sociological Methodology 1970. San Francisco: Jossey-Bass.

Duncan, Otis Dudley, Raymond P. Cuzzort and Beverly Duncan

1961 Statistical Geography--Problems in Analyzing Areal Data. Glencoe, Illinois: The Free Press of Glencoe.

Duncan, Otis Dudley, David L. Featherman and Beverly Duncan

1972 Socioeconomic Background and Achievement, New York: Seminar Press.

Farkas, George

1974 "Specification, residuals and contextual effects," Sociological Methods and Research 2 (November): 333-63.

Fennessey, James

1968 "The general linear model: a new perspective on some familiar topics." American Journal of Sociology 74 (July):1-27.

Gilbert, J. P. and Frederick Mosteller

1972 "The urgent need for experimentation. In F. Mosteller and D. P. Moynihan (eds.), on Equality of Educational Opportunity. New York: Vintage.

Goldberger, Arthur S.

1972 "Selection bias in evaluating treatment effects: some formal illustrations." Discussion paper 123-72. Institute for Research on Poverty. University of Wisconsin, Madison, Wisconsin.

Gordon, Robert A.

1968 "Issues in multiple regression." American Journal of Sociology 73(March): 592-616.

Hannan, Michael T. and Leigh Burstein

1974 "Estimation from grouped observations," American Sociological Review 39 (June):374-392.

Hannan, Michael T., John H. Freeman, and John W. Meyer

1976 "Specification of models for organizational effectiveness," American Sociological Review 41 (February):136-143.

Hauser, Robert M.

1968 Family, School, and Neighborhood Factors in Educational Performances in a Metropolitan School System. Unpublished doctoral dissertation, Department of Sociology, University of Michigan, Ann Arbor.

Hauser, Robert M.

1969 "Schools and the stratification process." *American Journal of Sociology* 74(May):587-611.

Hauser, Robert M.

1970 "Context and consex: a cautionary tale." *American Journal of Sociology* 75(January):645-664.

Hauser, Robert M.

1970 "Hauser replies," *American Journal of Sociology* 76 (November):517-520.

Hauser, Robert M.

1971 *Socioeconomic Background and Educational Performance.* Rose Monograph Series. Washington D. C.: American Sociological Association.

Hauser, Robert M.

1974 "Contextual analysis revisited," *Sociological Methods and Research* 2 (February):365-75.

Hauser, Robert M. and David L. Featherman

1976 "Equality of schooling: trends and prospects." *Sociology of Education* 49 (April):99-120.

Hauser, Robert M., William H. Sewell and Duane F. Alwin

1976 "High school effects on achievement," In W. H. Sewell, R. M. Hauser, and D. L. Featherman (eds.), *Schooling and Achievement in American Society*, New York: Academic Press

Heise, David R.

1972 "Employing nominal variables, induced variables, and block variables in path analysis." *Sociological Methods and Research* 1(November):147-173.

Heyns, Barbara

1974 "Social Selection and stratification within schools," _American Journal of Sociology_ 79 (May):1434-1451.

Jencks, Christopher and Marsha Brown

1975 "The effects of high schools on their students," _Harvard Educational Review_ 45 (August): 273-324.

Jencks, Christopher, Marshall Smith, Henry Acland, Mary Jo Bane, David Cohen, Herbert Gintis, Barbara Heyns, Stephan Michelson

1972 Inequality. A Reassessment of the Effect of Family and _Schooling in America_. New York: Basic Books.

Johnston, J. J.

1972 _Econometric Methods_. New York: McGraw-Hill.

Lazarsfeld, Paul F. and Herbert Menzel

1969 "On the relation between individual and collective properties." In amitai Etzioni (ed.) _Reader on Complex Organizations_. Pp. 499-516. New York: Holt, Rinehart and Winston.

Lord, F. M. and M. R. Novick

1968 _Statistical Theories of Mental Test Scores_. Reading, Mass.: Addison-Wesley.

Mayeske, G. W., T. Okada, A. E. Beaton, Jr., W. M. Cohen, and C. E. Wisler

1973 A study of the achievement of our nation's students. Washington, D.C.: U.S. Department of Health, Education and Welfare.

Meyer, John W.

1970 "High school effects on college intentions." _American Journal of Sociology_ 76 (July):59-70.

Michael, John A.

1961 "High school climate and plans for entering college," _Public Opinion Quarterly_ 25 (Winter): 585-95.

Mosteller, Frederick and Daniel P. Moynihan (eds.)

1972 On Equality of Educational Opportunity. New York: Random House.

McDill, Edward L., Leo C. Rigsby and E. D. Meyers, Jr.

1969 "Educational climates of high schools: their effects and sources." American Journal of Sociology 74 (May): 567-86.

Nelson, Joel I.

1972 "High school context and college plans: the impact of social structure on aspirations." American Sociological Review 37 (April): 143-148.

Riley, Matilda W.

1964 "Sources and types of sociological data," Pp. 978-1026 in R. E. L. Faris (ed.), Handbook of Modern Sociology, Chicago: Rand McNally & Company.

Rogoff, Natalie

1961 "Local structure and educational selection," Pp. 241-51 in A. H. Halsey, J. Floud, and C. A. Anderson (eds.), Education, Economy and Society. New York: The Free Press.

Schuessler, Karl

1971 Analyzing Social Data. Boston: Houghton-Mifflin.

Sewell, William H. and J. Michael Armer

1966 "Neighborhood context and college plans." American Sociological Review 31(April): 159-168.

Slocum, Walter L.

1966 Educational Aspirations and Expectations of Students in Rural Washington High Schools. Department of Rural Sociology. Washington State University, Pullman.

Tannenbaum, Arnold S. and Jerald G. Bachman

1964 "Structural versus individual effects." American Journal of Sociology 69 (May): 585-95.

Turner, Ralph H.

1964 The Social Context of Ambition. San Francisco: Chandler Publishing Company.

Werts, Charles and Robert L. Linn

1971 "Analyzing school effects: ANCOVA with a fallible covariate." Educational and Psychological Measurement 31(Winter): 95-104.

Werts, Charles E. and Robert L. Linn

1971 "Considerations when making inferences within the analysis of covariance model." Educational and Psychological Measurement 31(Summer): 407-416.

Wilson, Alan B.

1959 "Residential segregation of social classes and aspirations of high school boys," American Sociological Review 24 (December): 836-45.

RESEARCH DESIGNS REQUIRED FOR CONTEXTUAL AND AGGREGATE STUDIES OF SCHOOL AND SCHOOLING EFFECTS*

John W. Meyer
Department of Sociology
Stanford University

October, 1976

---

* Prepared for the Conference on Methodology for Aggregating Data in Educational Research, Stanford University, October, 1976. This paper spells out research design implications of some ideas discussed in my paper "The Effects of Education as an Institution" (Vasquez Associates & Department of Sociology, Stanford University, August, 1976). Copies of this paper are available.

The two most important findings of American sociological research on school effects create a paradox requiring explanation. Finding 1: When backgrounds and initial qualities of students are held constant, American schools of any given type do not produce very different effects. This is true across a considerable range of dependent variables. Finding 2: The length of time an individual has attended schools is the single most important background variable independently affecting a large number of status and social-psychological properties of individuals as they reach adult society. These two findings produce inconsistent interpretive discussions observing (1) that schools have few effects, and (2) that they have many large effects. Researchers who compare schools reach the first conclusion: Those who compare the schooled with the unschooled reach the second. Schooling is in one sense very much a dichotomous variable, not a continuous one -- on any resource dimension which has been examined, a good school produces only slightly better results than a bad one, but any school is much better than no school at all.

The explanation of this anomalous set of findings will obviously lie in discovering important school properties affecting educational outcomes which characterize practically all schools, but which do not vary much among schools. Three relevant, and partly distinct, theories have been suggested:

> Theory 1: Teaching is very important in affecting educational outcomes, but teaching quality does not vary much among schools. Most teachers are effective in comparison with no teacher at all. Beyond this, good teaching is almost randomly assigned to schools, which employ teachers on credentialling criteria little related to teaching skill.
>
> Theory 2: School effects arise mostly from the time and attention the schools require students to devote to particular topics. American schools are relatively homogeneous on this factor.
>
> Theory 3: School effects are created (a) by the institutional certification rules which give graduates enormous advantages over other persons (e.g., in entering college, in entering the professions, and so on), and (b) by the awareness students and their social networks contain about these rules. American high schools (and American colleges) have little variation in their certificational status, and are thus relatively homogeneous in their effects.

Research Design Implications

Theories 1 and 2 suggest that we should look for contextual effects, not by comparing schools, but by comparing micro-contexts within (and between) schools which vary on (1) teaching effectiveness and/or (2) the time and attention devoted to topics relevant to the dependent variables of interest. Such research designs are fairly well understood and do not require discussion here. Theory 3 has more complex research design implications.

The idea that schools have their effects in part because of their certificational power is supported by evidence that where the certification status of schools or programs varies, outcomes tend to vary too:

1. Type of School: In countries in which high schools are distinguished in clearly-defined types which differ in formal access to the universities, these types show quite different effects. Graduates of English grammar schools have higher expectations and actual social status outcomes that graduates of secondary modern schools of similar background and intelligence. (The available studies, however, do not control quite all the individual properties of known relevance.) In less well controlled data, Polish gymnasium graduates end up in dramatically higher status occupations than graduates of vocational secondary schools of similar class background. More well-controlled and longitudinal studies of educational systems which are certificationally tracked are needed.

In this country, longitudinal studies starting at the beginning of high school are called for. Holding constant backgrounds and initial abilities and intentions, these studies could compare our public high schools with (a) vocational high schools, and (b) private college preparatory schools, since these are the only types of high schools we have which differ sharply on certificational status.

2. Curriculum: Within high schools, college preparatory and vocational curricula clearly differ in certificational status. And several studies suggest that they differ in their effects on initially similar students. We need longitudinal studies of the high school years, however, to more carefully control initial self-selection to these curricula..

At the college level, a number of studies show differences -- though not always large ones -- in the effects of different curricula or majors. Factors other than certificational status, however, are usually invoked as explanatory.

3. Schooling itself, in every country in which its effects have been investigated, dramatically alters student orientations and life chances. As we noted above, contrasts between people in an organization the Ministry of Education considers a school and those not in such an organization show large differences. These effects may result from the credentialled status of the schools, but as noted above there are clearly other interpretations.

4. Finally, complex comparative research designs could study directly the effects of credentialling rules themselves. If one could compare the effects of similar schools on similar students in countries which differed in their credentialling and allocation rules, Theory 3 could be tested rather directly. Do elite secondary schools have more effects on the orientations and future careers of given students in countries in which these schools have more rights to confer entry into elite universities? The most feasible research of this kind would involve comparing the effects of professional training programs in countries which differ in the extent to which these training programs -- and the professions themselves -- are licensed to carry great social authority.

The obvious conclusion here is that if schools derive many of their effects from their credentialling status in society, we should find schools which vary on this attribute. In order to understand the effects of an independent variable, we have to find research designs which contain variance on the variable. We have enough studies which look for variance in school organizational structure. We have too few studies which look for variance in micro-climates within schools. And we have almost no studies which maximize variance in the institutional rules which define the meaning and status of schools and which link them (and their graduates) to society.

Aggregate Effects

School effects data are not simply evaluated as results in their own right. They are almost always interpreted as suggesting aggregate effects on society as a whole. This is highly suspect: more complex research designs are needed to show that the "effects" of schools on individuals can be aggregated to show societal effects. The best known example of this kind of reasoning is the long literature in the economics of education. This literature starts with the conventional evidence that schooled persons make much more money than less schooled persons of similar background. It is then argued that these income advantages can more or less be aggregated (with some minor corrections) to show the societal benefits of education. Since the educated in fact tend to have higher incomes, this line of inference has been used around the world to justify educational expansions in the national interest. The sociological literature on school effects tends to produce the same kind of inference, but with more imagistic language.

School effects on students cannot simply be combined to produce estimates of overall effects because these effects may not be independent of other, unstudied, consequences of the schools. The schools, that is, may affect not only students, but other members of society: and their effects on these other persons may be intimately linked with (either directly or inversely) effects on students. Two cases have been discussed:

Case 1: The Negative Effects of Schools on Non-Students: The schools may affect students in part because they credential these students with favorable life-chances. But consider the effect of these credentials on non-students. The life-chances -- and presumably the orientations and anticipations -- of such persons are negatively affected by the number of educated people around them and by the degree to which these educated people are credentialled. The rewards and resources of the social system may be fixed in sum, and the expansion of those going to the educated may be balanced by losses for the less educated. Inasmuch as educational effects are the effects of credentialling rules, the most dramatic differences between the educated and the uneducated may be found even though no aggregate effect on society as a whole exists at all.

Many studies now show the operate of these kinds of "frog-pond" effects: But we have no way of knowing -- without more comparative research -- how important they may be in considering the societal effects of education. Our point here is methodological: In order to discover the aggregate effects of education we must have comparative data. We cannot simply compare schooled and unschooled persons within one society. We must compare each class of individuals with similar people in a less schooled society to see whether the gains of the educated (a) are a net controbution of education or (b) occur at the expense of the uneducated.

Case 2: The Effects of Schools on Members of Society Generally: The educational system is a network of rituals and social classifications which penetrates society in many ways. Its construction and expansion may greatly alter the ways all sorts of societal members and elites think and take action.

The creation, in other words, of a huge social structure classifying and legitimating many types of personnel and knowledge is bound to have consequences for ordinary persons, political elites, economic managers, and so on, with their own educational levels held constant. Persons of any given educational level, in other words, will be likely to think and act differently depending on whether they are surrounded by the authorized people and authoritative knowledge created by the modern educational system.

Almost all the sociological thinking along these lines suggests that the effects of the schools created in this way will be "positive" -- that is, that the creation of the schools in general infects the general population with the same beliefs and orientations as it does the students. In principle, however, some effects along these lines could be negative -- as persons become generally incapacitated to hold values or undertake activities now reserved for specialists created by the educational system.

Aggregate Effects: Research Design Implications

The intellectual considerations noted above make it clear why one cannot easily infer aggregate school effects from a controlled comparison of students and nonstudents within a system. (1) Such studies miss the fact that student-nonstudent differences might have been created by an education-based worsening of the condition of the latter. In this respect these studies overestimate the aggregate effects of education. (2) But such studies also miss the effects of education as an institution on members of society in general -- students and nonstudents alike. In this respect these studies underestimate the effects of education.

Proper studies of the aggregate effects of education must include variance on the obvious independent variable, comparing societal instances in which education is more and less widely spread and deeply institutionalized. We have considered several ways in which variations in education may affect the attitudes and life-chances of individuals when their own educations are held constant. If such institutional effects exist, then institutional variations in education must be included as independent variance.

But studies must also include variation in the full range of dependent variables which might add up to an aggregate effect. There are two ways in which this can be done:

1. We can do longitudinal aggregate studies, showing the effects of variations in education on such societal properties as economic growth and political development. In this way we get summary estimates of the aggregate effects of education through all the different process by which they might occur. We learn nothing from such studies, however, about how the effects in fact occur: for instance, whether they arise from effects on students or effects on others.

2. We can try to do comparative contextual studies, examining the effects of national variations in education on the properties of students, nonstudents, and other social personnel (e.g.,

ordinary members, elites). Research of this kind might be very expensive. It also might tell us a good deal about how education works in society.

Conclusion: In order to study the aggregate effects of education, we need to be sure that both aggregate aspects of the independent variables and of the dependent variables are included in research designs. (1) As an independent variable, we must include variance in education as an institution, not simply variance in the education of individuals. This clearly requires comparative research. (2) We must include in our dependent variables measures of the likely effects. At the minimum, we need to consider (a) direct effects on students, (b) negative effects on nonstudents through the reduction in opportunities available to them, and (c) effects on all types of members of society arising from the existence and legitimation of new kinds of personnel and information in society.

Unraveling Classroom Effects:

A Multilevel Feedback Model of Contextual Effects

Alice A. Young,
Graduate School of Industrial Administration
Carnegie-Mellon University

Lutz Erbring
Institute for Social Research
University of Michigan

Paper prepared for the Conference on Methodology for Aggregating Data in Educational Research, October 22-23, 1976, Palo Alto, California

Introduction

The theoretical and statistical status of "contextual effects" has been a subject of continuing debate by sociologists, political scientists, educational researchers, and members of related disciplines. This controversy has persisted for a quarter of a century, though the intensity of debate has varied. Despite many conceptual and methodological arguments concerning the need to distinguish genuine results of group composition from methodological artifacts, however, there is little consensus about how to assess the impact of group-level processes. (For example, see Cronbach, 1975). Either the question of whether contextual effects exist is knotty indeed, or else there must be something wrong with the way the question is usually conceived.

In principle, most researchers would agree that the more accurately a model assumed in estimation reflects the underlying process, the more confidence can be placed in the corresponding empirical results. More precise specifications of the generating mechanisms for contextual effects, therefore, should go a long way toward preventing spurious interpretations of statistical results.

This paper focuses on a contextual model which is based on the notion of endogenous feedback and is applicable to the study of educational aspirations and achievement, a subject which received fresh impetus with the publication of the "Coleman Report." The proposed model offers a new way to evaluate the impact of students' interaction with each other, whether in the classroom or the school. After presenting examples of hypotheses for which the technique is useful, we discuss the type of estimation appropriate to the model. Finally we compare the proposed strategy of analysis to some traditional styles of assessing the impact of contextual effects.

I. A Model of Student Interaction

Suppose we wish to model the determination of levels of academic achievement for each student in a given set of classrooms. Academic achievement in this instance might involve performance in a particular subject or completion of a sequence of training. There are two determinants of performance we wish to consider. First, we hypothesize that each student's intellectual ability as measured, e.g., by intelligence quotients, affects his or her academic performance. Second, students' contacts with each other affect each individual's level of achievement. We postulate that the higher the educational attainments of the students with whom an individual interacts, the higher will be that individual's own attainments.

Clearly the assumptions just presented are an oversimplified view of reality. In actual applications of the proposed technique, a more complete set of specifications should obviously be employed. However, we focus upon this simple case because it represents the two general types of mechanisms which underlie the determination of student academic performance. One component of that process represents the effects of exogenous factors such as an individual's intelligence, family background, etc., or a classroom's teacher quality, curriculum content, etc. The other component of the process represents the effects of student social interaction. This endogenous component is the result of students comparing themselves, competing with each other, and learning from one another, whether within a classroom or across a school.

Let the data consist of observations for all students in J classroom with I_j students per class. The academic achievement of the i^{th} student in the j^{th} classroom will be labeled y_{ij}, while that student's measured intellectual ability will be denoted x_{ij}. To represent our assumptions about achievement we then employ the following model:[1]

$$y_{ij} = \alpha\left(\sum_{\ell=1}^{J} \sum_{k=1}^{I_j} w_{(k\ell)(ij)} y_{k\ell}\right) + \beta_1 x_{ij} + u_{ij}\,, \qquad (1)$$

$$(i = 1,\ldots,I_j;\ j = 1,\ldots,\ J)$$

where:

α, β, are the regression coefficients corresponding to the terms

$$\left(\sum_{k=1}^{I_j} w_{(k\ell)(ij)} y_{k_j}\right) \text{ and } x_{ij}$$

respectively. α is a feedback coefficient representing the effect of interaction among students, while β is an impact coefficient representing the effect of a student's individual ability. $w_{(k\ell)(ij)}$ is a weight pertaining to the extent of interaction between student i of class j and student k of class ℓ. For two students who do not interact, $w_{(k\ell)(ij)}=0$.

u_{ij} is an independently distributed error term, distributed $N(0, \sigma^2)$.

The general form of the model in equation (1) would be (assuming from here on that $I_1 = I_2 = \ldots = I_j$ for convenience):

$$\underline{y} = \alpha \underline{Wy} + \underline{X\beta} + \underline{u}, \qquad (2)$$

where:

$\underline{y}$ is an I·J by 1 vector of outcome scores, y_{ij}'s,

$\underline{X}$ is an I·J by P matrix of exogenous variables, which could include a constant term,

α is the feedback coefficient,

β is a P by 1 vector of regression coefficients for the exogenous variables,

$\underline{W}$ is an I·J by I·J matrix of fixed numbers representing the social structure through which the I·J units of observation are connected with each other, and

$\underline{u}$ is an I·J by 1 vector of disturbances, u_{ij}'s, with $\underline{u} \sim N(0, \sigma^2 \underline{I})$. ($\underline{I}$ is the identity matrix.)

The structure of the weight matrix, $\underline{W}$, is an integral part of the model. In general, $\underline{W}$ represents exogenous information about physical or social contiguity among members of the collectivity. Thus, it mirrors an essential aspect of the social structure characterizing a given student population, and it defines the conduit for a variety of substantive processes which involve transactions among individuals -- persuasion, contagion, competition, cooperation, imitation, etc. Clearly, $\underline{W}$ has to be specified a priori, either on the basis of data about the incidence of relevant pairwise contacts among students or, in the absence of more specific information, by resorting to simplifying assumptions about patterns of student interaction.

The definition of contiguity may vary somewhat depending on the type of process assumed to generate the contextual feedback effects. Thus, in the case of communication or emulation processes, contiguity may have to be based on amount of interaction specific to each pair of students; whereas in the case of comparison or competition processes, contiguity may be treated as fixed and equal for all pairs of students in a given classroom and zero for all pairs of students from different classrooms. The latter case of equal within-class contiguity and zero between-class contiguity may, of course, also be a plausible assumption about the probability of face-to-face interaction in the absence of further information.

In conjunction with the appropriate contiguity matrix $\underline{W}$, the feedback rate α is of critical importance from a substantive point of view. It provides a measure of the extent to which the outcome variable for each individual, e.g., academic performance, is affected by contiguous "spill-over" from other members of the group. Specifically, if the rows of $\underline{W}$ are normalized to sum to unity, e.g., by expressing each individual's amount of contact with others in terms of proportions, then the metric of the dependent variable will be preserved by the endogenous feedback term. As a result, α can be interpreted as the degree to which individual outcomes y_{ij} are determined by contextual as opposed to exogenous mechanisms.

In order to appreciate the role of the endogenous feedback component of the model in the contextual determination of individual outcomes, it is instructive to rearrange the terms in equation (2) to give:

$$(\underline{I}-\alpha\underline{W})\underline{y} = \underline{X}\underline{\beta} + \underline{u} \tag{3}$$

Equation (3) expresses the dependent variable, net of feedback effects, as a function of the exogenous variables and the disturbances. The reduced form of (3) is

$$\underline{y} = (\underline{I}-\alpha\underline{W})^{-1}\underline{X} + (\underline{I}-\alpha\underline{W})^{-1}\underline{u} \qquad (4)$$

But the inverse matrix in equation (4) is the limit of a convergent infinite series of the form

$$(\underline{I}-\alpha\underline{W})^{-1} = \underline{I} + \alpha\underline{W} + (\alpha\underline{W})^2 + \ldots$$

provided that $-1 < \alpha < 1$. In that case, equation (4) can be written as

$$\underline{y} = \underline{X}\underline{\beta} + (\alpha\underline{W})\underline{X}\underline{\beta} + (\alpha\underline{W})^2\underline{X}\underline{\beta} + \ldots + (\underline{I}-\alpha\underline{W})^{-1}\underline{u} \qquad (5)$$

Formally, equation (5) represents the equilibrium of the feedback structure of the model. Substantively, equation (5) illustrates that the final scores on the endogenous variables incorporate the results of cumulative filtering of the effects of exogenous variables, beginning with their direct impact on each individual member of the group; and continuing with their indirect effects due to the impact of exogenous variables on those with whom (s)he interacts, on those with whom his (her) first-order interaction partners interact; and so on.[2] Thus, the admissible values of α must fall between 0 and 1 since it would be rather implausible to assume that indirect effects due to the impact of exogenous variables on other members of the group become strengthened rather than weakened with increasing "distance" from a given student or that they alternate in sign depending on whether the number of intervening links is even or odd.[3]

II. Estimation of the Model

Ordinary least squares estimation of α and β from equation (2) would produce inconsistent estimates of these parameters. Since a linear combination of observations for the dependent variable appears on the right hand side of (2), there is feedback between the regressors and the error term in the model. We can picture this violation of the standard OLS assumptions more clearly from an examination of Figure 1. This illustration shows the feedback process involving the respective academic achievement levels of students in a class j which consists of three individuals who interact with one another with relative frequencies w_{ij}, but have no significant contact outside their class. Thus, the model is a simultaneous equation structure, and the endogenous variables must be purged of feedback effects before consistent estimates of α and of $\underline{\beta}$ can be obtained.

In order to address the estimation problem, it is useful to proceed from the rearranged form of the basic model given by equation (3) above, which expresses the dependent variable, net of feedback effects, as a function of the exogenous variables and disturbances. If an estimate of α could be secured separately and if the existence of the inverse matrix $(\underline{I}-\alpha\underline{W})^{-1}$ was assured, the remaining parameters could be estimated conditional on a fixed value of α, from the reduced form in equation (4), by employing the appropriate transformation to the exogenous variables. But the resulting estimates would still not be consistent, since the reduced-form transformation would necessarily destroy the independence property for the transformed disturbances,

$$\begin{aligned} E[(\underline{I}-\alpha\underline{W})^{-1}\underline{uu}'(\underline{I}-\alpha\underline{W})^{-1}{}'] &= (\underline{I}-\alpha\underline{W})^{-1}E[\underline{uu}'](\underline{I}-\alpha\underline{W})^{-1} \\ &= (\sigma^2\underline{I})(\underline{I}-\alpha\underline{W})^{-2} \end{aligned}$$

which is a non-diagonal matrix.

A maximum likelihood estimator which avoids the problems associated with reduced-form estimation in this case has been proposed

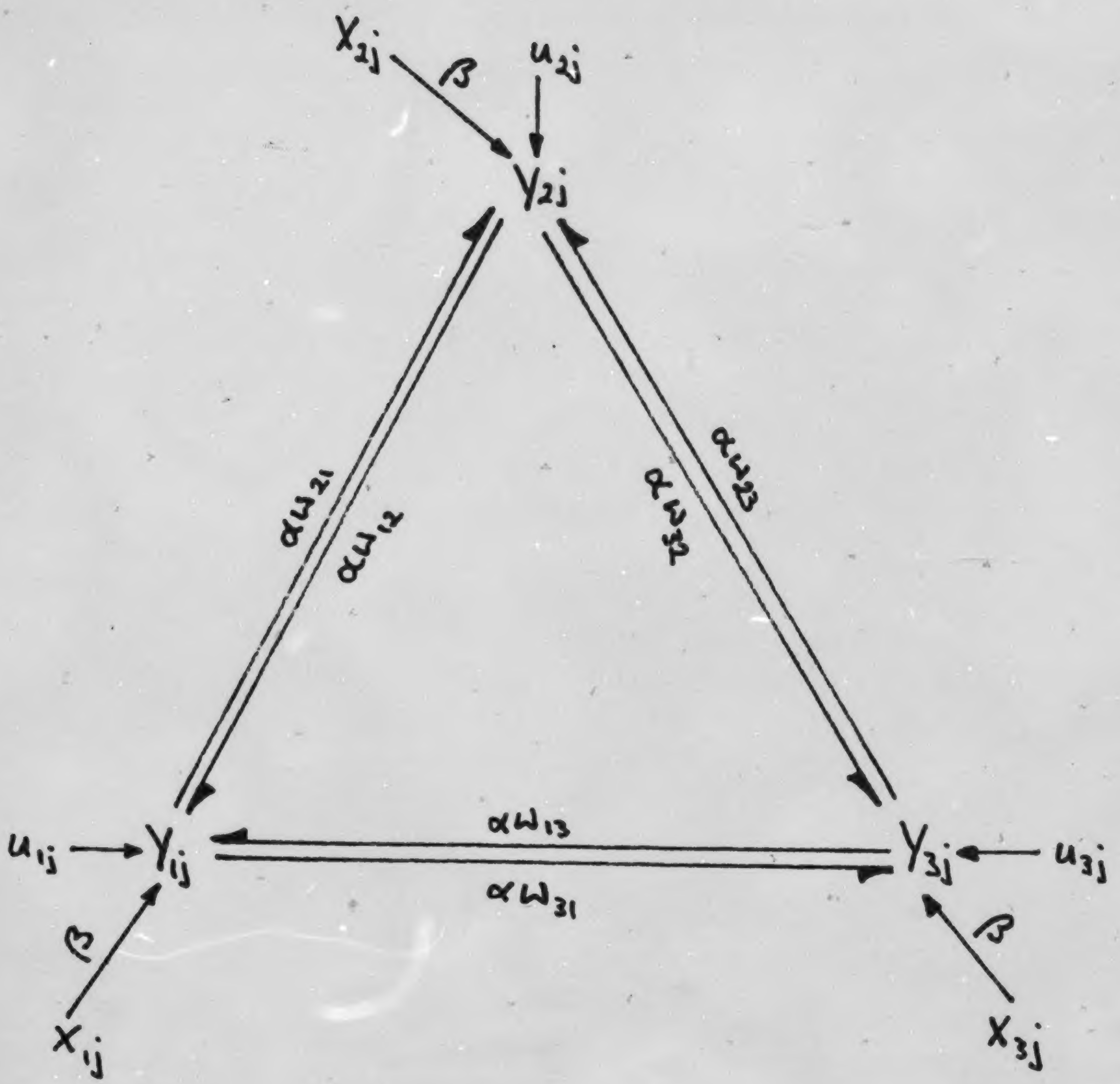

X2j
β
u2j
Y2j
αω21
αω12
αω23
αω32
u1j
Y1j
αω13
αω31
Y3j
u3j
β
β
X1j
X3j

Figure 1

Contextual Feedback Model for a 3-Student Classroom recently by Lebanon et al. (1975), based directly on the rearranged form of the structural model in (3) above, which contains the original disturbances. Since these disturbances were assumed to be identically and independently distributed, it follows that the likelihood of $\underline{y}^* = (\underline{I}-\alpha\underline{W})\underline{y}$ is a function of the density of $\underline{u}$. Finally, the observed $\underline{y}$ are linear transforms of the $\underline{y}^*$, which contain the unknown parameter α. Thus, assuming $f(\underline{u}) = N[\underline{0}, \sigma^2\underline{I}]$, and using the Jacobian of the transformation $\underline{y} = (I-\alpha W)^{-1}\underline{y}^*$, the resulting likelihood function of $\underline{y}$ is

$$
\begin{aligned}
L[\underline{y}] &= L[(\underline{I}-\alpha\underline{W})\underline{y}]\,|\mathrm{Det}(\underline{I}-\alpha\underline{W})| \qquad (6)\\
&= (2\pi\sigma^2)^{-N/2}\exp[-(1/2\sigma^2)[((\underline{I}-\alpha\underline{W})\underline{y} - \underline{X\beta})'((\underline{I}-\alpha\underline{W})\underline{y} - \underline{X\beta})]\\
&\quad |\mathrm{Det}\ (\underline{I}-\alpha\underline{W})|
\end{aligned}
$$

yielding the log-likelihood function

$$
\begin{aligned}
L^*[\alpha,\beta,\sigma^2;\underline{y}] &= N/2\ \ln(2\pi) - N/2\ \ln(\sigma^2) - \qquad (7)\\
&\quad 1/2\sigma^2[((\underline{I}-\alpha\underline{W})\underline{y}-\underline{X\beta})'((\underline{I}-\alpha\underline{W})\underline{y}-\underline{X\beta})] + \ln|\mathrm{Det}\ (\underline{I}-\alpha\underline{W})|
\end{aligned}
$$

where N=I·J as defined earlier.

Maximizing L* with respect to $\underline{\beta}$ and σ^2 yields

$$\hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'(\underline{I}-\alpha\underline{W})\underline{y} \qquad (8)$$

and, using (8)

$$
\begin{aligned}
\hat{\sigma}^2 &= [\underline{y}'(\underline{I}-\alpha\underline{W})'(\underline{I}-\underline{X}(\underline{X}'\underline{X})^{-1}\underline{X}')(\underline{I}-\alpha\underline{W})\underline{y}]/N \qquad (9)\\
&= [Q]/N
\end{aligned}
$$

where Q is the error sum of squares resulting from $\hat{\underline{\beta}}$ in (8). These estimators are conditional upon α, which must be estimated by iterative grid search based on its concentrated likelihood function. The concentrated likelihood of α, using (8) and (9), is given by

$$L^*[\alpha;\underline{y}] = -N/2\ \ln\ (2\pi/N) - N/2 - N/2[\ln(Q)] + \ln|\mathrm{Det}\ (\underline{I}-\alpha\underline{W})|$$

Thus the solution involves finding the smallest value of $Q^{N/2}/|\mathrm{Det}(\underline{I}-\alpha\underline{W})|$.

III. Alternative Models of Contextual Effects

The relationship between our contextual feedback specification and other, more conventional formulations of contextual effects can best be illustrated by tracing the consequences of incorrectly omitting the endogenous feedback component from the model. Thus if the true model is

$$\underline{y} = \alpha\underline{Wy} + \underline{X\beta} + \underline{u}.$$

but the analysis is based instead on a model

$$\underline{y} = \underline{X\beta} + \underline{e}$$

where $\underline{e} = \alpha\underline{Wy} + \underline{u}$.

then the OLS estimates will be biased. The expectation of these incorrectly specified OLS estimates would be:

$$E(\hat{\beta}) = \beta + E((\underline{X}'\underline{X})^{-1}\underline{X}'\alpha\underline{Wy}). \tag{10}$$

Obviously the effects of the specification bias in (10) depend on the form of $\underline{W}$, recalling that there are a number of different ways to specify $\underline{W}$. We only consider the special case in which pairs of student interactions are equally weighted within classrooms, while no significant interaction occurs across classrooms. This specification is most readily tractable and corresponds to the typical design for conventional discussions of contextual effects. Thus, we assume that, for a given school, $\underline{W}$ is a block-diagonal matrix containing blocks of $(\underline{\ell\ell}'/I)$ on the diagonal (where $\underline{\ell}$ is a vector of unities). For the case of $W = \ell\ell'/I$, the basic model becomes

$$\underline{y} = \alpha\bar{\underline{y}} + \underline{\beta X} + \underline{u}$$

where

$$\bar{y} = \underline{I} \otimes (\underline{\ell\ell}'/I)\underline{y}$$

is a vector of group means. Accordingly, the expectation of the estimator reduces to

$$E(\underline{\hat{\beta}}) = \underline{\beta} + E[(\underline{X}'\underline{X})^{-1}\underline{X}'\underline{\bar{y}}\alpha] \tag{11}$$

If the model without the endogenous feedback term were to be estimated from a "partitioned" (within-group) design separately for each of the j classrooms, then the bias would remain confined to an intercept shift (since $\bar{y}$ is constant within a given group).[4] In fact, the vector of $\bar{y}$ would be indistinguishable from the constant term of the original model, and an estimate of α could not be obtained even if the theoretically appropriate feedback term were included for estimation (i.e., the data matrix would be singular).

If instead the model without the endogenous feedback term were to be estimated from a "total" (within and between group) design, the bias would extend to the coefficient estimates for the exogenous variables. The reason for the latter effect becomes readily apparent by recognizing, first, that the regression of group means $\bar{y}$ on the exogenous variables (i.e., $(\underline{X}'X)^{-1}\underline{X}'\underline{\bar{y}}$ in (11) above) controls the bias, the second, that that bias is necessarily nonzero whenever the regression of individual scores, $\underline{y}$, on the same exogenous variables (i.e., $(\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}$ in general) is nonzero to begin with.[5]

There are several alternative strategies which may be employed in order to extract the bias due to omission of the endogenous feedback component. One of these would be to use a pooled design with a more complex error structure which acknowledges the existence of unspecified group-level effects, e.g., in the form of group-specific dummy variables. While this removes the bias from the coefficient estimates for the exogenous variables, it does not purge bias from the model altogether; it merely reallocates it to the group intercept terms (thus restoring the results obtained from partitioned estimation). Furthermore, it precludes any possibility of distinguishing this particular source of bias from the effects of other possible specification errors.

The second strategy is to compensate for the omission of endogenous feedback would be to introduce the group means of one or more of the exogenous variables as additional regressors in the pooled design (recalling that the omitted feedback term is necessarily correlated at the group level with each of the exogenous variables in the model). This is essentially equivalent to the addition of group dummy variables and accomplishes the same result: It removes the bias from the exogenous coefficient estimates and allocates it to the coefficients associated with the group means (which yield, in effect, partial intercept shifts for the groups). Unlike the use of dummy variables, however, it invites substantive interpretations in terms of "group-level effects" which, under the circumstances, are evidently spurious.

The third strategy, and obviously the one which we would advocate, would be to remove the bias by proper specification of the model, i.e., by allowing for the possibility of endogenous feedback. Unlike either of the other approaches, it represents a substantive process whose validity is based on explicit a priori considerations rather than merely on statistical adjustments for group-level errors. The question of whether a particular outcome such as student performance can be associated with a specifiable process predicated on physical or social contiguity at the individual level must be resolved on theoretical grounds. Furthermore, the answer clearly does not depend in any way on the particular kinds of exogenous variables involved -- individual abilities, group treatments, social background, or whatever. Indeed, if the notion of contextual effects is to signify substantive rather than statistical phenomena, then the assumption that behavior begets behavior (mediated through a given social structure) has much to recommend itself. If should not, of course, be used indiscriminately; where contiguity is not involved, explanations of "group-level" effects, should they occur, must probably be sought elsewhere.

Conclusion

In sum, the specification of endogenous feedback allows -- or compels -- us to make explicit the mechanisms believed responsible for "contextual effects" in the determination of outcomes such as student achievement. The model reflects the premise that behavioral outcomes directly impinge upon behavioral outcomes by virtue of actual contact between individuals. Conversely, the model avoids the need to postulate action at a distance through reified "group norms." The specific interpretation of context reflected in W, of course, ultimately depends on the type of social process assumed on *a priori* theoretical grounds.

FOOTNOTES

1. The inspiration of our model comes from Mitchell (1969). Mitchell uses a similar idea to treat the effects of spatial contiguity in explaining the take-over of local governmental units by guerrilla insurgence in Southeast Asia.

2. The parameter α is analogous to the feedback rate in a dynamic model, but here α is weighted by a two-dimensional structure of social contiguity rather than a one-dimensional structure of temporal contiguity.

3. For values of α less than 0, the signs on odd-numbered powers of $(\alpha\underline{W})$ in the expansion in (5) would be negative, while the signs on even-numbered powers of $(\alpha\underline{W})$ would be positive.

4. This derivation is based on the following algebraic manipulations, assuming $\underline{W}=\ell\ell'/I$:

Let: $\underline{X}_1 = [\ell\ \tilde{\underline{X}}_1]$, where ℓ is an I by 1 vector of 1's, and $\tilde{\underline{X}}_1$ is the I by (P-1) matrix of observations on the regular regressors for classroom 1; and let y_1 be the I by 1 vector of observations on the dependent variable for classroom 1.

Then: $(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_1'\alpha\underline{W}\underline{y}_1 = \alpha(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_1'\underline{W}\underline{y}_1$

$$= \alpha\begin{bmatrix} \ell'\ell & \ell'\tilde{\underline{X}}_1 \\ \tilde{\underline{X}}_1'\ell & \tilde{\underline{X}}_1'\tilde{\underline{X}}_1 \end{bmatrix}^{-1} \begin{bmatrix} \ell'\underline{y}_1 \\ \tilde{\underline{X}}_1'\frac{\ell\ell'}{I}\underline{y}_1 \end{bmatrix}$$

$$= \alpha\begin{bmatrix} (\frac{1}{I})\ell'y_1 \\ \underline{0} \end{bmatrix}$$

The last step was based on the formula for the inverse of of a partitioned matrix given in Theil (1971, p. 18).

5. To extend the analysis in Footnote 4 to a group of classrooms, we could let $\underline{W}=(\underline{I}\otimes\ell\ell'/I)$. For this latter definition of $\underline{W}$;

$(\underline{X}'\underline{X})^{-1}\underline{X}'\alpha\underline{W}\underline{y} = \alpha(\underline{X}'\underline{X})^{-1}\underline{X}'\underline{W}\underline{y}$

$$= \alpha\begin{bmatrix} \ell'\ell & \ell'\tilde{\underline{X}} \\ \tilde{\underline{X}}'\ell & \tilde{\underline{X}}'\tilde{\underline{X}} \end{bmatrix}^{-1} \begin{bmatrix} I\sum_{j=1}^{J}\bar{y}_j \\ \sum_{j=1}^{J}\tilde{\underline{X}}_j'\ell\bar{y}_j \end{bmatrix}$$

(where $\bar{y}_j$ is the mean of the values of the dependent variable for the j^{th} classroom, and $\underset{\sim}{X}_j$ is an I by (P-1) matrix of observations for the regular exogenous variables for the j^{th} classroom).

Upon expanding the expression after the last "=" sign, we found that for $\underline{W} = (I \otimes \ell\ell'/I)$ the terms for both the constant term and for the slope coefficients for the regular regressors are non-zero.

BIBLIOGRAPHY

Cronbach, Lee J.
1976 "Research on Classrooms and Schools: Formulation of Questions, Design, and Analysis." Stanford, California: Stanford Evaluation Consortium, School of Education.

Lebanon, A.; T.W. McGuire; H. Rosenthal; and M. Lee
1976 "Cross Sectional Dependence: The Econometrics of the Huk Rebellion Reconsidered," to appear in Political Methodology.

Mitchell, E.J.
1969 "Some Econometrics of the Huk Rebellion," American Political Science Review (63), pp. 1159-1171.

Theil, Henri
1971 Principles of Econometrics. New York: John Wiley and Sons.

Independence and Selecting the Appropriate Unit of Analysis (or the Appropriate Analytic Model)

Linda K. Glendening
National Institute of Education

And

Andrew C. Porter
Michigan State University

Paper presented at the Conference on Methodology for Aggregating Data in Educational Research, sponsored by a contract from the National Institute of Education (NIE-C-74-0123), at Stanford University, Stanford, California, October 23-24, 1976.

We are interested in the class of designs which attempt to test for effects of an independent variable (some set of interventions) on a dependent variable. Further, the interventions are of a type which are hoped to affect individual performance but which are delivered to groups of individuals, e.g., classrooms. If observations on the dependent variable are collected at the individual level, the analysis can either include the grouping variable or not. In much of the literature, the choice between the two options is argued on the grounds of whether or not observations at the individual level are independent of each other. If yes, then one might not include the grouping variable; if no, then presumably there is independence at the group level. More specifically then, our work is an attempt to investigate the effect of violations of the assumption of independence on tests of the hypothesis of no treatment effects on a dependent variable; or alternatively, the effect of testing that same hypothesis with an incorrectly specified model. It is our contention that the literature arguing choice of unit of analysis, vis-a-vis concern for the assumption of independence, can profit from seeing that the concern is directly analogous to proper model specification.

Briefly, our study contained three interrelated components:

- The first component entailed deriving an operational definition of independence of responses.
- The second component used that definition to study the consequences of using correlated units of analysis, or in other words, to study the effect of violating the assumption of independence within an ANOVA context.

- And the third component looked at the questionable procedure of doing a preliminary test to select a proper unit of analysis to use in a subsequent test of treatment effects.

In addition to summarizing our research outlined above, we briefly consider dependence within the ANCOVA model.

Our Operational Definition of Independence

The motivation for our research came from the already existing definitions of independence. To us these definitions seemed inadequate as they were conflicting in the suggested causes of dependence and deficient in the assessment of dependence. And being able to operationalize the assumption of independence is essential for studying the consequences of violating that assumption.

For the most part, our definition of independence is based on the work of Peckham, Glass and Hopkins, who have suggested two threats to independence of responses. They call the first threat an additive effect. This affect raises or lowers the group mean by affecting each score within the group similarly. This effect influences the variability between group means, while doing nothing to the within classroom variability. For example, a differential teacher effect on classrooms could cause just such an additive effect. On the other hand, the second threat, what they call a proportional effect, increases or decreases the variability of responses within the group but does nothing to affect the between class variability. As an example, a differential teacher effect within classrooms could cause this proportional effect. These two threats and their affect on the between class and within class variabilities provided the basis for our definition of independence. We propose that independence be operationally defined as that condition where the expected mean square between groups equals that within groups. More generally, this definition states that disaggregate units can be treated as independent units whenever the variance of the aggregate units can be predicted from the grouping size and the variance of the disaggregate units.

Given the above definition, two types of dependency are possible, positive and negative. Positive dependence was defined as the expected mean square between groups being greater than the expected mean square within groups, or similarly by the intraclass correlation coefficient being greater than zero. Negative dependence was defined by the expected mean square within groups being larger than the expected mean square between groups, or similarly by the intraclass correlation coefficient being less than zero. Either type of dependence can be caused by an additive effect, a proportional effect or simply nonrandom assignment to groups. At this point it may be worth noting that random assignment of subjects to groups does not necessarily erase the possibility of positive or negative dependence occuring, but that nonrandom assignment almost insures such problems.

Importantly, the proposed operational definition of independence seems to be a useful one within the typical educational framework of hierarchically-nested designs where students are nested within classrooms, classrooms nested within schools, etc. Further this definition has the advantage of being readily estimable.

Correlated Units of Analysis

Armed with an operational definition of independence we studied the consequences of violating that definition. We first did it analytically to determine the existence and direction of any effects and then empirically to estimate the magnitude of effects. For the simulation part of the study, we chose relatively small degrees of positive and negative dependence to work with, as it is with small to intermediate degrees of dependence that the effects of dependence seem most nebulose. The actual degrees of positive dependence, measured by the amount of student variation that is accounted for by the classroom variation, that we used fit well with that found in the Follow Through data analyzed by Haney (1974).

The effects of violating the assumption of independence were studied within the context of a balanced, two-level hierarchically-nested design, having subjects nested within groups and groups nested within treatments. Two ANOVA models were used to analyze the data fitting this general description. One model (the so-called "never pool" model) used group or classroom as the unit of analysis and was an always correct model in this study; while the other model (the "always pool" model) used student. With the assumptions underlying this study, namely that classrooms or the aggregate units were independent of each other and were thus correct units of analysis, the question of interest was: Under conditions of dependence, what happens when student or the disaggregate unit is used as the unit of analysis?

Given the independence case, student and classroom are both appropriate units of analysis. The "always pool" model with student as the analytic unit is preferred, however, as it has more degrees of freedom error, which means more power. As an example of the effect increasing the degrees of freedom error had on power where the size of the treatment effect was .4 within class standard deviations, given the number of classrooms per treatment was 5 and the number of students per class was 20, the power discrepancy at α equal .05 between using class as the unit of analysis and student as the unit of analysis was .111, favoring the model using student as the unit.

Given positive dependence (as we define it), the proper unit is the group unit. Using the disaggregate unit caused the pooled error term for the "always pool" test to be too small and thus the test of treatment effects was too liberal. And our empirical research showed this liberalness to be of such size to be of practical importance. As expected, the power of the "always pool" test was spuriously high and not interesting to look at.

Given negative dependence, the correct unit is once again the grouped unit as we assumed the group units to be independent and using the ungrouped unit caused the "always pool" F test to be a conservative test. This conservativeness spuriously reduced the empirical power to such an extent that the advantage of using the "always pool" test in the first place (more d.f. error) was cancelled out.

What do the results suggest? They indicate quite clearly that the F test is not robust to violation of the assumption of independence, even with small degrees of dependence. This conclusion should be kept in mind both when designing and analyzing experiments as well as when interpreting results of studies which have not followed the advice from this investigation. For example, studies of treatment effects using traditional teaching methods where students are in classrooms and learn as a group may be coming up with too many significant results if student is used as the unit of analysis. Or if instruction somehow is making students within a class extremely heterogeneous and student is used as the unit of analysis, studies of treatment effects may be coming up with too few significant results. In addition, given the results of this study, researchers designing new teacher studies, for example, may well think about selecting more schools and fewer teachers within any one school to include in their study than selecting few schools and more teachers within any one school.

Effects of the Preliminary Test

Because our research showed it to be very important (in inferencial situations) to have independent units of analysis, one might next ask, Is it feasible to use a conditional testing procedure by first doing an initial test of independence, testing the equality of the expected mean squares between and within classes, and then on the basis of that test choosing unit of analysis (student or classroom) for the primary test of treatment effects? This study showed clearly, both analytically and empirically, that the answer to this question is no.

Only where ungrouped units were independent did this conditional or "sometimes pool" procedure turn out to be effective, but this procedure was not as effective under this case as the "always pool" test for increasing the power.

Given positive dependence, the conditional testing procedure turned out to be too liberal and consequently had spuriously high power.

Given negative dependence, the conditional test was somewhat conservative and generally had less power than the "never pool" always correct F-test, even though it had more d.f. error.

In summary then, this study showed that, in a hierarchically-nested design with one outcome measure per subject, as a general rule of thumb a preliminary test of independence should not be used to choose a unit of analysis to test for treatment differences. If individual units are independent, the "always pool" F test is best. If individual units are not independent, the "never pool" F test is best. The problem is, however, the researcher is never in the position before hand of actually knowing whether or not responses of subjects nested within groups are, in fact, independent responses.

ANCOVA and Dependence

For ANCOVA, the assumption is that the disaggregate scores adjusted by the covariate (X) are independent of each other both within and across treatments. In our earlier research this assumption would be met when the expected mean square for classes equals the expected mean square for students within classes. For ANCOVA, these would be adjusted mean squares.

The question of interest is: Does the use of ANCOVA do anything to help the "use of the wrong unit" problem? That is, can ANCOVA correct for dependence in student within class data, allowing student to be used as the unit of analysis? Or alternatively, can the model be correctly specified using covariates as opposed to a grouping variable?

When considering the utility of ANCOVA to solve the dependence problem, it seems necessary to make some distinctions as to when and how dependence arises; distinctions that were not necessary in the earlier consideration of ANOVA. For example, in the ANOVA case, for statistical purposes at least, it isn't important to know where or when the dependence occurred. One need only recognize that it was present in the data. This doesn't seem to be the case with ANCOVA designs. That is, in ANCOVA situations, it seems to be important to know where dependence in the posttest variable (Y) comes from. For example, is the dependence reflected in the covariable (X)? If not, did it enter in during the treatment phase? Does any dependence in the X variable at pretest time change over time?

Three ANCOVA models can be used to study our problem. The first model (Model A) has two sources of variation, treatments (T) and classes within treatments (C:T) and is an "always correct" model to use when classrooms are independent of each other. Model B also has two sources of variation, treatments (T) and students within treatments (S:T). Model C has three sources of variation, treatments (T), classes within treatments (C:T) and students within classes (S:CT).

Let's take the positive dependence case where, on both X and Y, the variation of student observations within the classroom is too small in relation to the variation of classroom means. This means that in Model C, the expected mean square for classrooms, on the unadjusted dependent variable Y, is greater than the expected mean square for students within classrooms. The trick in handling dependence is to somehow get these two mean squares equal and equal to the expected mean square treatment under the null.

The expected mean squares for C:T and S:CT in Model C can be made equal by either reducing the $E(MS_{C:T})$ or increasing the $E(MS_{S:CT})$. We will consider reducing the $E(MS_{C:T})$. The two expected mean square formulas would be equal if the term $\sigma^2_{C:T}$, which is greater than zero under positive dependence, could somehow be removed from the $E(MS_{C:T})$ in Model C. This is comparable to the problem of removing covariate differences in quasi-experiments. Acknowledging the great debate about the utility of ANCOVA for such purposes, pretend for the moment that ANCOVA makes the appropriate adjustment and so makes $\sigma^2_{C:T}$ equal to 0. If this is indeed the case, the adjusted expected mean squares, T, C:T and S:CT in Model C are now all equal under the null hypothesis. And thus the adjusted mean squares C:T and S:CT can be combined to form a pooled error term in testing for treatment effects. Even though using the pooled adjusted mean squares C:T and S:CT is like using students as the unit of analysis, we don't believe the pooled and adjusted error term just described equals the adjusted subjects within treatments (S:T) error term in Model B. Model C uses the pooled within class regression coefficient ($\beta_{S:CT}$) to do the adjusting; while Model B uses the pooled within treatment regression coefficient ($\beta_{S:T}$) and given positive dependence these two coefficients should not be equal.

It seems that the procedure outlined above can be effective if (1) the variance of X and Y are equal, which means that X and Y have equal degrees of dependence, and (2) the correlation of X and Y within classrooms ($r_{S:CT}$) is equal to one.

As stated previously, in the ANCOVA situation, it is important to know where the dependence affected the data. If the design calls for random assignment of students to classes and classes to treatments, there should be no dependence at pretest or pre-treatment time. However, if there is differential teacher effectiveness, for example, during the treatment period, there will be dependence at posttest time. In this case, since there are no initial differences to remove, ANCOVA cannot solve the dependency problem as $\sigma^2_{C:T}$ will remain greater than zero in Model C and using the "wrong unit", student, should cause problems similar in direction, at least, to those in the ANOVA situation.

The real point to be made, in the situation described above (where ANCOVA should correct for positive dependence by making the $\sigma^2_{C:T}$ term equal to zero), is which model (Model A with class as the unit or Model C with student as the unit) should be used to analyze the data. As one attempt to determine this, one can think of the error variance in ANCOVA as primarily a product of two factors: one, the variance of the dependent variable and the other, one minus the correlation between the dependent variable and covariate squared. The variance of the dependent variable should be smaller in Model C than A since ANCOVA removes $\sigma^2_{C:T}$, which is greater than zero given positive dependence, from the error variance in Model C. On the other hand, the correlation should be larger in Model A than Model C. If the differences between models for the two factors counterbalance, the preference is given to Model C because of greater degrees of freedom error. This, however, makes the strong assumption that the covariate actually results in a $\sigma^2_{C:T}$ equal to zero.

Small Sample Results on Estimation from Grouped Observations

Michael T. Hannan
Department of Sociology
Stanford University
and
Alice A. Young
Graduate School of
Industrial Administration
Carnegie-Mellon University

August, 1976

*The research reported here was supported by NIE research contract C-75-0123. Lutz Erbring, François Nielsen and Leigh Burstein made valuable comments on an earlier draft.

Introduction

Warnings concerning the pitfalls of estimation from grouped observations have been sounded in virtually all the social science fields over very long periods of time. A number of recent works have clarified the mechanisms by which grouping affects the behavior of least squares estimators (Blalock 1964; Hannan 1971; Feige and Watts 1972; Hanushek et al. 1974; Hannan and Burstein 1974). Despite the long history of concern and the recent upsurge of interest in the problem, a great deal of current research practice appears virtually unaffected. This gap between methodological developments and research practice reflects, at least in part, the relative imprecision of the guidelines concerning the magitude of the errors involved in analyzing grouped observations. Although we have made progress in determining the conditions under which grouping observations leads to bias, we know very little about the likely magnitudes of bias. If researchers are to confront the tradeoffs involved in replacing easily available aggregated data with more appropriate disaggregated data, they must be given at least rough estimates of the likely costs of the errors involved in using the grouped data.

There is a second sense in which the methodological literature fails to meet the needs of researchers. Virtually all the attempts to explicate the mechanisms producing bias in grouped estimators have focused either explicitly or implicitly on two variable models. Since such models are hardly ever used in research and because it is so often the case that issues arise in multivariate models that do not pertain to bivariate models, it is important to extend the study of grouping effects to more complex models.

There are formidable obstacles (identified below) to an analytic treatment of the full complexity of the grouping problem. We utilize Monte Carlo simulations to begin to remedy the two limitations just noted. The use of simulated data enables us to calculate precisely the bias and efficiency costs of a variety of types of grouping. In addition, we investigate the effects of grouping in multiple regression models. We also consider how the effects of grouping depend on model specification. By so doing, we provide a concrete illustration of available analytical results in addition to providing a number of new results. These findings are addressed both to situations in which the researcher can control the way in which data is to be aggregated and to situations in which the analyst must treat the data in whatever form of grouping has already been provided.

1. The Model

Assume one has a sample of observations on N individuals, with each case generated by the following model:

$$y_i = \beta_o + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \epsilon_i$$

or,[1] in matrix notation: (1)

$$\underline{y} = \underline{X}\,\underline{\beta} + \underline{\epsilon}$$

Further assume the following[2] hold:

$$\begin{array}{rll} \text{i.} & E\,(\,\underline{\varepsilon}/\underline{X}\,) & = 0 \\ \text{ii.} & E\,(\,\underline{\varepsilon\varepsilon}'/\underline{X}\,) & = \mathrm{var}(\varepsilon_i)\cdot I \\ \text{iii.} & E\,(\,\underline{\varepsilon}'\underline{X}/\underline{X}\,) & = 0 \end{array} \tag{2}$$

For instance, (1) might represent the regression of an individual student's achievement on family background, ethnicity, classroom size, etc.

2. *The Estimators*

We wish to compare the behavior of two least squares estimators of $\underline{\beta}$. The first, called the *ungrouped estimator*, employs the N sample observations:

$$\underline{\hat{\beta}} = (\,\underline{X}'\underline{X}\,)^{-1}\underline{X}'\underline{y} \tag{3}$$

Under the conditions stated in (2), $\underline{\hat{\beta}}$ is a best linear unbiased estimator of $\underline{\beta}$.

The second least squares estimator uses m groupings or averages of the N observations, with $m < N$. Following Prais and Aitchison (1954), we introduce a grouping matrix $\underline{G}$, of size (m by N), that transforms[3] the N original observations into m grouped observations. Using this notation we can write the *grouped estimator* as:

$$\underline{\hat{\beta}}(g) = (\,\underline{X}'\underline{G}'\underline{GX}\,)^{-1}\underline{X}'\underline{G}'\underline{Gy} \tag{4}$$

Obviously the behavior of the grouped estimator depends on the nature of the grouping matrix. If, as Prais and Aitchison (1954) assumed, the elements of $\underline{G}$ are fixed constants, then $\underline{G}$ is uncorrelated by definition with either the disturbance or any of the observed variables. Then the grouped estimator is unbiased but inefficient relative to the ungrouped estimator. But in almost all cases of interest the elements of the grouping matrix are stochastic, as in the cases discussed by Blalock (1964) and Hannan (1971). In realistic situations the grouping rule will usually be related to one or more of the variables under analysis. Such is the case when observations are placed in groups on the basis of location in social structure (e.g., classroom, place of residence, etc.).

The recent literature on grouped estimators has investigated a number of simple stochastic groupings. To evaluate the properties of the grouped estimator, we substitute in (4) for $\underline{y}$:

$$\begin{aligned} \hat{\underline{\beta}}(g) &= (\underline{X}'\underline{G}'\underline{GX})^{-1}\underline{X}'\underline{G}'\underline{G}(\underline{X}\,\underline{\beta} + \underline{\varepsilon}) \\ &= \underline{\beta} + (\underline{X}'\underline{G}'\underline{GX})^{-1}\underline{X}'\underline{G}'\underline{G\varepsilon} \end{aligned} \qquad (5)$$

The conditions under which $\hat{\underline{\beta}}(g)$ may be expected to be unbiased if $\hat{\underline{\beta}}$ is unbiased are parallel to (2) above. In particular, for $\hat{\underline{\beta}}(g)$ to be unbiased the critical feature of the grouping method is that, equivalent to iii) in (2) above:

$$\underline{E}\;(\underline{\varepsilon}'\underline{G}'\underline{GX} \mid \underline{GX}) = 0 \qquad (6)$$

That is, there must be no stochastic relationship between the grouped values of $\underline{\varepsilon}$ and any variable in $\underline{GX}$.

Any grouping that retains the zero covariance between the disturbance and the regressors will yield unbiased grouped estimation. Two types of grouping that have this property have been widely discussed. The first is random grouping. In this case it follows from elementary sampling theory that (6) will hold and that $E[\hat{\underline{\beta}}(g)] = \underline{\beta}$. The second case involves grouping matrices whose elements are functions of the regressors (and are thus uncorrelated with the disturbance). The best known case of this type is what we have termed "grouping by X", i.e., grouping that maximizes variation in one or more vectors of $\underline{GX}$.

A number of situations have been identified where the independence of disturbance and regressors breaks down. This is the case for "grouping by Y" (i.e., grouping that maximizes variation in $\underline{Gy}$) as shown by Blalock (1964).

The mechanisms leading to bias and inefficiency in grouped estimators have been explicated clearly only for the simplest models (two-variable regression models). This is due no doubt to the difficulty of obtaining analytic results (e.g., taking expectations of (5)) when the grouping matrix is stochastic. We employ Monte Carlo simulations to overcome this difficulty. By simulating a variety of grouping schemes we obtain precise information on the effects of grouping on the behavior of least squares estimators. And, we show how the results from the simple models can be extended to more complex cases.

Besides presenting results on the effects of grouping in well specified models, we also investigate grouping effects for models that omit relevant causal variables. This analysis extends the results of Hannan and Burstein (1974) and Hannan et al. (1975) on the manner in which grouping bias interacts with specification bias from omitted variables.

3. The Monte Carlo Simulation

Previous empirical attempts to assess the effects of grouping upon estimation quality, as in the school effects research (e.g., Hannan and Burstein, 1974; Hannan *et al.*, 1975), have indicated that, at least under some conditions of grouping, the resulting estimates have marked increases in bias, variance, or both. These results displayed sizeable discrepancies in the estimates from grouped data as compared with ungrouped data. However, there are several deficiencies in such empirical comparisons. Since the exact mechanisms generating the empirical observations were unknown, it is possible that the ungrouped estimates were based upon incorrectly specified models. Moreover, it was impossible to observe how the stochastic relationships between the unobserved true error terms and the variables under consideration were altered by the various grouping rules.

The use of a Monte Carlo simulation provides some remedies for these problems. Since the correct regression model is known for simulated data, the exact bias and variance of both grouped and ungrouped estimators can be established. In addition, the moments of the true error term--and the covariance of the error term with the regressors--may be calculated for each grouping method. The moments involving the error term can be used to explicate the effects of aggregating data.

A. Design of the Simulation

We designed a Monte Carlo simulation for a population characterized by

$$Y = 1X_1 + 1X_2 + \delta u \ .$$

X_1 and X_2 were both N(0,1), u was N(0,1) and distributed independently of X_1 and X_2, and δ was a parameter set to vary R^2 for the equation.[4] We varied R^2 (at .3 and .7) and the correlation between X_1 and X_2, ρ_{12}, (at 0, .25, .5, and .75). For each of the eight parameter combinations of R^2 and ρ_{12}, 100 samples of size 500 were generated.

We then arranged the 500 individual cases into 50 equal-sized groups by each of four grouping methods: random, by values of X_1, by values of X_2, and by values of Y. For example, in grouping by X_1 we ordered observations in decreasing order by X_1 values and placed the first 10 observations in one group, the next 10 in the next group, etc. Then we calculated the means of X_1,X_2, and Y for each group. This procedure was repeated for each method of grouping to yield all of the grouped data used in the following analyses. The variances and covariances of ungrouped and grouped observations on X_1, X_2, Y, and ϵ are presented in Table 1.

[Table 1 here]

In the discussion which follows let $\mathrm{var}(X_1)'$ and $\mathrm{var}(X_1)$ denote the variances of the ungrouped and grouped observations of X_1, respectively, and so on for the other variances and covariances.

B. Analysis Framework

There are two phases to our analysis. The first concerns the effects of grouping on estimates from a properly specified model. In this phase we compare the ungrouped and grouped ordinary least squares estimators:

$$\hat{\underline{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}, \text{ where } \underline{X} = (\underline{X}_1, \underline{X}_2) \qquad (7)$$

and:

$$\hat{\underline{\beta}}(g) = \begin{pmatrix} \hat{\beta}_1(g) \\ \hat{\beta}_2(g) \end{pmatrix} = (\underline{X}'\underline{G}'\underline{GX})^{-1}\underline{X}'\underline{G}'\underline{Gy}\ , \qquad (8)$$

where (8) is estimated for four types of grouping: (1) random grouping, (2) grouping by X_1, (3) grouping by X_2, and (4) grouping by Y.

The second phase of analysis concerns the interactions of omitted variables bias with grouping bias. Therefore we consider two biased ordinary least squares estimators:

$$\hat{\beta}_{1(m)} = (\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_1'\underline{y} \qquad (9)$$

and

$$\hat{\beta}_{1(mg)} = (\underline{X}_1'\underline{G}'\underline{GX}_1)^{-1}\underline{X}_1'\underline{G}'\underline{Gy} \qquad (10)$$

Throughout the discussion which follows, two summary statistics will be used to describe the quality of the estimators of β_1 and β_2. The average bias indicates how close estimated values of a parameter, say β_1, come to the true value, while the variance measures the amount of dispersal estimates display around their mean value. Thus, the average bias of $\hat{\beta}_1$, based on 100 samples of simulated data, is calculated as:

$$\text{Average Bias}\left(\hat{\beta}_1\right) = \frac{1}{100}\sum_{i=1}^{100}\left(\hat{\beta}_{1i} - \beta_1\right).$$

Likewise, the variance of $\hat{\beta}_1$ is estimated as:

$$\text{Variance}\left(\hat{\beta}_1\right) = \frac{1}{100}\sum_{i=1}^{100}\left[\hat{\beta}_{1i} - \left(\frac{1}{100}\sum_{i=1}^{100}\hat{\beta}_{1i}\right)\right]^2$$

4. Results for a Properly Specified Model

Table 1 confirms that the moments in the randomly grouped data are all approximately one-tenth the values of the ungrouped counterparts. Therefore the results on bias and efficiency are very simple (cf. Cramer 1964). There is of course no grouping bias. The efficiency of the grouped estimators is slightly less than one-tenth that of the ungrouped estimators. As Cramer (1964) points out, regression coefficients estimated from N observations aggregated randomly into m groups have about the same efficiency as estimates based on just m randomly chosen observations. (A simulation study is not necessary to produce these results on bias and efficiency. We show these findings for use as baselines in analyzing the results for the more analytically untractable conditions.)

Case 2. Grouping by X_1

When observations have been grouped by X_1, we note from Table 2 that $\hat{\beta}_1(g)$ and $\hat{\beta}_2(g)$ are unbiased. This result can be easily interpreted from an examination of Table 1. The covariances between X_1 and ε and between X_2 and ε remain approximately zero in the aggregated data. Thus the analytical prediction of unbiasedness for estimates of β_1 and β_2 under grouping by X_1 is confirmed.

[Table 2 here]

The simulation study allows us to extend previously known results for estimator variance from the one-regressor case. Cramer (1964) demonstrates that when there is only one independent variable, grouping which maximizes between-group variation will the most efficient. This means that grouping by X_1 will be much more efficient than random grouping, although no method of aggregation can exceed the efficiency of ungrouped data for the model in (1) though (4).

Several patterns displayed in the efficiency results of Table 3 require further exploration. It does seem intuitively reasonable that, since grouping by X_1 maximizes the variation in X_1, the variance of $\hat{\beta}_1(g)$ is always less for grouping by X_1 than for random grouping. However, both the strong relationships between ρ_{12} and the variance of $\hat{\beta}_1(g)$ as well

[Table 3 here]

as the lack of association between ρ_{12} and the variance of $\hat{\beta}_2(g)$ necessitate a closer examination. It is not immediately apparent why the relative efficiency of $\hat{\beta}_1(g)$ is one when X_1 and X_2 are uncorrelated but declines sharply as that correlation departs from zero. Further, the relative efficiency of $\hat{\beta}_2(g)$ is approximately equal to that for random grouping, almost independent of ρ_{12}. This is surprising since grouping by X_1 is not random with repect to X_2 when ρ_{12} is nonzero.

To understand these results, it is helpful to note the following: (1) grouping by X_1 does not systematically alter any covariance involving X_1; (2) almost no variance in X_1 is lost in this type of grouping, i.e., $\mathrm{var}(X_1) \cong \mathrm{var}(\bar{X}_1)$; and (3) the variance in all other variables is lowered by this grouping. To see how these results produce the findings on efficiency reported in Table 3 we need an expression for the variances of the estimators

$$\mathrm{var}(\hat{\beta}_1) = \mathrm{var}(\epsilon)/(1-r_{12}^2)\mathrm{var}(X_1) \tag{11}$$

$$\mathrm{var}(\hat{\beta}_2) = \mathrm{var}(\epsilon)/(1-r_{12}^2)\mathrm{var}(X_2) \tag{12}$$

and (letting $\bar{r}_{12}^2$ denote the correlation between the grouped values of X_1 and X_2):

$$\mathrm{var}[\hat{\beta}_1(g)] = \mathrm{var}(\epsilon)/(1-\bar{r}_{12}^2)\mathrm{var}(\bar{X}_1) \tag{13}$$

$$\cong \mathrm{var}(\epsilon)/(1-\bar{r}_{12}^2)\mathrm{var}(X_1) \tag{14}$$

$$\mathrm{var}[\hat{\beta}_2(g)] \cong \mathrm{var}(\epsilon)/(1-\bar{r}_{12}^2)\mathrm{var}(X_2) \tag{15}$$

The approximate equality in (14) holds in our data because almost no variance in X_1 is lost with grouping-see Table 1. With this approximation, the relative variance of ungrouped to grouped estimators of β_1 has the following simple form:

$$\text{Rel. Var.}\left(\hat{\beta}_1/\hat{\beta}_1(g)\right) \simeq \frac{1-\bar{r}_{12}^2}{1-r_{12}^2} \tag{16}$$

And only one of the three quantities that enters $\bar{r}_{12}^2$ is affected by grouping: $\text{var}(\bar{X}_2)$.

Now we use the results above to explain how the expression in (16) changes when we start at $\rho_{12} = 0$ and steadily add to the value of ρ_{12}. Since $\text{var}(\bar{X}_2)$ is less than $\text{var}(X_2)$ and all other quantities are approximately unchanged, $\bar{r}_{12}^2 > r_{12}^2$ --except when $\rho_{12} = 0$, in which case $\bar{r}_{12}^2 \cong 0$. This latter result explains why the relative variance of ungrouped to grouped estimators of β_1 is roughly equal to one when $\rho_{12} = 0$. As ρ_{12} increases the following effects occur in the quantities that enter $\bar{r}_{12}^2$ and r_{12}^2: (1) $\text{var}(X_1)$ and $\text{var}(\bar{X}_1)$ are approximately equal to one; (2) $\text{cov}(X_1,X_2)$ and $\text{cov}(\bar{X}_1,\bar{X}_2)$ are both approximately equal to ρ_{12}; (3) $\text{var}(X_2)$ approximately equals one, while $\text{var}(\bar{X}_2)$ decreases; and (4) $\text{cov}(X_1,X_2)$ (and $\text{cov}(\bar{X}_1,\bar{X}_2)$) increase at a much slower rate that that at which $\text{var}(\bar{X}_2)$ decreases. These four effects combine to make the relative variance $\hat{\beta}_1$ to $\hat{\beta}_1(g)$ decline sharply as ρ_{12} rises from zero. This difference in relative efficiency is most severe for moderate and high levels of ρ_{12} (.5 and above).

In the case of estimators of β_2,

$$\text{Rel. Var. } \left(\hat{\beta}_2/\hat{\beta}_2(g)\right) \cong \frac{\left(1-\bar{r}_{12}^{\,2}\right)}{\left(1-r_{12}^{\,2}\right)} \cdot \frac{\text{var}(\bar{X}_2)}{\text{var}(X_2)} \qquad (17)$$

The two ratios on the right-hand side of (17) respond in the opposite direction to increases in ρ_{12} increases the first ratio on the right hand side of (17) decreases as we just discussed. But as ρ_{12} increases, grouping by X_1 is decreasingly random with respect to X_2 and the second ratio on the right (the ratio of the two variances) increases. The two tendencies must offset each other so that over all values of ρ_{12} the effeciency of $\hat{\beta}_2(g)$ is quite similar to that of randomly grouped estimators, as we noted in considering Table 3.

Case 3. Grouping by Y

As shown in Table 1, grouping by the dependent variable leads to substantial covariances between the dependent variable and the disturbances. With simulated date we have the advantage of knowing the true values of the error terms, so that we may directly observe the changes in $\text{cov}(X_1, e)$ and $\text{cov}(X_2, e)$ brought about by grouping by Y. In Table 2 two facts are noteworthy. The bias introduced by grouping by the outcome is considerable, ranging from 34% to 196% of the true parameter value of the parameter. Second, the magnitude of such bias varies considerably with the R^2 of the equation (cf. Shively 1969; Hannan 1971). The lower the R^2 the greater the upward bias introduced by grouping by Y.

From Table 3 we find that the estimators produced by grouping by the dependent variable have relatively large variances. In no sense is the large bias offset by small variability of the estimators. Thus, large distortions from the true values of the regression coefficients may be expected in this case.

5. Results for an Improperly Specified Model

The problem gets more complicated and more interesting when the model being estimated is misspecified. We consider the case in which X_2 is incorrectly omitted from the model as in (8) above. That is, the analyst estimates from a two variable model containing only X_1 and Y when the regression in (7) also containing X_2 is actually the correct form.

Clearly the omission of X_2 biases the ungrouped estimator except in the case in which X_1 and X_2 are uncorrelated in the population. The expected value of the ungrouped estimator can be written as follows:

$$E\left(\hat{\beta}_1(m)\right) = \beta_1 + \beta_2 \left(\frac{\text{cov}\ (X_1, X_2)}{\text{var}\ (X_1)}\right)$$

Since in our simulation, β_2, var (X_1), and var (X_2) are all equal to unity, the specification bias (the second term on the right-hand side of (18)) is equal to the correlation of X_1 and X_2. Table 4 reports the estimated bias for the ungrouped estimator. The results conform closely to what is expected from (18).

[Table 4 here]

Case 1. Random Grouping

As long as the grouped values of X_1 and ϵ are uncorrelated, the bias in $\hat{\beta}_1$(mg) will be equal to:

$$\beta_2 \cdot \left(\frac{\text{cov }(\bar{X}_1, \bar{X}_2)}{\text{var }(\bar{X}_1)} \right) \tag{19}$$

Now all of the moments in the randomly grouped data are equal to one-tenth of their ungrouped counterparts. Thus, for the case of random grouping the specification bias in (19) will be equal to the specification bias in (18). This result is demonstrated in Table 4.

Case 2. Grouping by X_1.

For grouping by the included regressor, X_1, the formula in (19) again expresses the bias in $\hat{\beta}_1$(mg), since the grouped values of X_1 and ϵ are uncorrelated. In order to evaluate (19) for the case of grouping by X_1 we turn to Table 1. In this case the grouping retains virtually all of the ungrouped variation in X_1 and does not alter the covariance of X_1 and X_2. As a consequence, we would expect the specification bias for grouping by X_1 to be roughly equal to that for the ungrouped misspecified model. We see that this analysis is confirmed in Table 4.

The results on efficiency are presented in Table 5. As we see, the grouped estimator is fully as efficient as the ungrouped estimator. So for this particular method of grouping the grouped and ungrouped estimators are nearly identical.

[Table 5 here]

Case 3. Grouping by X_2.

Quite different results hold when the data are grouped by the omitted variable. However, the formula in (19) again applies because with grouping by X_2 the grouped values of X_1 and ϵ are still uncorrelated. The covariance of X_1 and X_2 is unaltered by grouping by X_2, but the variance of X_1 is attenuated (see Table 1). Grouping that maximizes variation in X_2 averages out some of the variation in X_1. The lower the correlation between X_1 and X_2, the greater the loss of variation in X_1. Thus, the lessened variation in X_1 will make the bias in $\hat{\beta}_1$(mg) much larger than for the ungrouped estimator. We see in Table 4 that the increase is very large. In the best case ($\rho_{12} = .75$) the percentage increase in the bias is 50%, in the worst ($\rho_{12} = .25$) it is approximately 140%.

The results for ($\rho_{12} = 0$) add an important clarification to our understanding of the conditions under which grouping produces bias. When the excluded variable is uncorrelated with the included variable, obviously there is no specification bias in the ungrouped estimator. Nor, in our data, is the grouped estimator biased. What we see is that grouping by a part of the disturbance that is uncorrelated with the regressor(s) does not produce any correlation of the regressor(s) with the disturbance and thus does not generate grouping bias. Thus, the presence of a stochastic relationship between the grouping criterion and the disturbance does not by itself produce bias under grouping.[6]

We pause here to compare the variances of $\hat{\beta}_1$(mg) in Table 5 for random grouping, grouping by X_1, and grouping by X_2. With each of these types of grouping the relative efficiency of the grouped estimators to the ungrouped estimators is the relative efficiency predicted analytically for the correctly specified one-regressor models (cf. Cramer 1964). That is, $\hat{\beta}_1$(mg) is about one-tenth as efficient as $\hat{\beta}_1$(m). Further, since almost no X_1 variation is lost in grouping by X_1, the estimates of β_1 under grouping by X_1 are about as efficient as the ungrouped estimates, across all values of ρ_{12}. Finally, the greater the amount of variation in X_1 preserved in grouping by X_2(i.e., the higher the value of ρ_{12}), the greater the relative efficiency of the estimates of β_1 under grouping by X_2. The latter two findings are in marked contrast with those for the correctly specified two-regressor estimations presented above. We recall that for the correctly specified two-regressor case the relative efficiency of $\hat{\beta}_1$(g) under grouping by X_1 varied across values of ρ_{12}, while the relative efficiency of $\hat{\beta}_1$(g) under grouping by X_2 remained constant across values of ρ_{12}.

Case 4. Grouping by Y.

The impact of grouping by the dependent variable in the misspecified model intensifies the specification bias. This result is important because there are no available small-sample analytical results which indicate the magnitude of the combined effects of aggregation and misspecification for this case. As shown in Table 4, the biases range from 142% to 428% of the value of the parameters being estimated. Like the results for the correctly specified estimations, the deleterious effects of grouping by Y are inversely related to the size of R^2.

The basic mechanism through which bias occurs in the case of grouping by the dependent variable is the large covariance between X_1 and the disturbances. This stochastic relationship between X_1 and the error term has two basic sources. Regardless of whether $(\rho_{12} = 0)$ in the ungrouped data, after grouping by Y the values of X_1 and X_2 will always be correlated in the aggregate data. Second, as we discussed for the correctly specified model, grouping by Y produces a correlation between the aggregated values of X_1 and ϵ, even when $(\mathrm{cov}\ (X_1,\epsilon) = 0)$ in the ungrouped observations.

Discussion

These small sample results on the bias that arises when data are grouped so as to maximize variation in particular variables in a model agree closely with Hannan and Burstein's (1974) large sample results. Moreover the new evidence allows us to assess the relative magnitudes of bias and efficiency across the various types of grouping.

Just as important is the demonstration that extension of existing results to more complex models is straightforward. Grouping "by" a variable does not systematically alter zero-order covariances involving the variable. It does, however, change variances of other variables. Using these results, we see that grouping by one of a set of regressors does not yield aggregation bias (even when the model is misspecified). But grouping by endogenous variables (e.g., the dependent variable) or an omitted causal variable that is correlated with one or more regressors does produce aggregation bias. This result on the interaction of aggregation bias and specification bias has not seen noted previously. Grouping magnifies specification bias present in ungrouped estimates (due to omitted variables) when the method of grouping observations preserves grouped variation in omitted variable(s).

The simulation study also provides several important results for the effects of grouping on efficiency. First, the large biases in estimates based on grouping by the dependent variable are in no way compensated for by decreased variances. Thus grouping by the dependent variable produces the double penalties of large estimator bias and substantial loss in efficiency. Further, the case of grouping by one of the regressors for the correctly specified model provides an interesting extension of results for the bivariate model. Relative to random grouping, grouping by an independent variable produces smaller estimator variance for the coefficient that is associated with the grouping criterion. However, the greater the covariance between the two independent variables, the less the relative increase in efficiency.

The variance of the coefficient estimate for the regressor not used as a grouping criterion is about the same size as the variance in the case of random grouping.

These results can be extended to simultaneous equations models and to at least some dynamic models (cf. Hannan _et al_. 1975). However, it is important to point out that they pertain to only one form of grouping: grouping by values of some variable in a model. Empirical researchers typically face situations that involving grouping rules more complex than this. In addition, the effects of measurement error have not yet been completely analyzed. This problem is certainly prevalent in a great deal of empirical research. The results presented here are intended as a benchmark for treating more complex situations. It is clear that work on other types of grouping and on measurement error deserves deeper study.

Table 1. Values of Sample Moments of Observations

		Variance (X_1)				Variance (X_2)			
R^2	ρ_{12}*	Ungrouped	Randomly Grouped	Grouped by X_1	Grouped by Y	Ungrouped	Randomly Grouped	Grouped by X_1	Grouped by Y
.7	.75	.988	.096	.983	.636	.996	.098	.600	.645
.7	.5	.989	.095	.984	.558	.999	.099	.318	.568
.7	.25	.991	.095	.986	.481	1.002	.099	.150	.494
.7	0	.998	.098	.993	.409	1.002	.099	.099	.416
.3	.75	.988	.096	.983	.328	.996	.098	.600	.332
.3	.5	.989	.095	.984	.294	.999	.099	.318	.299
.3	.25	.991	.095	.986	.263	1.002	.099	.150	.269
.3	0	.998	.098	.993	.227	1.002	.099	.099	.236

		Variance (ϵ)				Covariance (X_1,X_2)			
R^2	ρ_{12}	Ungrouped	Randomly Grouped	Grouped by X_1	Grouped by Y	Ungrouped	Randomly Grouped	Grouped by X_1	Grouped by Y
.7	.75	1.501	.147	.143	.555	.742	.072	.738	.616
.7	.5	1.286	.126	.125	.476	.493	.048	.491	.514
.7	.25	1.072	.105	.101	.399	.245	.023	.244	.413
.7	0	.858	.084	.083	.315	-.002	.001	-.002	.312
.3	.75	8.171	.798	.778	5.950	.742	.072	.738	.306
.3	.5	7.004	.684	.679	5.095	.493	.048	.491	.246
.3	.25	5.836	.570	.550	4.249	.245	.024	.244	.191
.3	0	4.669	.456	.454	3.387	-.002	.001	-.002	.131

* Each entry based on 100 samples.

** The values of R^2 and of ρ_{12} given are those used to create the original 500-observation data sets.

Table 1. Values of Sample Moments of Observations, cont.

		Covariance (X_1,ε)				Covariance (X_2,ε)			
R^2	ρ_{12}	Ungrouped	Randomly Grouped	Grouped by X_1	Grouped by Y	Ungrouped	Randomly Grouped	Grouped by X_1	Grouped by Y
.7	.75	-.000	-.000	-.000	.469	.005	.000	-.002	.474
.7	.5	-.001	-.000	-.001	.401	.006	.000	-.002	.408
.7	.25	-.002	-.000	-.001	.334	.006	.000	-.001	.338
.7	0	-.003	-.000	-.003	.267	.005	.000	-.001	.273
.3	.75	-.000	-.000	-.001	1.087	.012	-.001	-.005	1.104
.3	.5	-.003	-.000	-.002	.933	.014	.001	.006	.953
.3	.25	-.004	-.000	-.003	.772	.014	.001	-.002	.795
.3	0	-.006	-.001	-.006	.625	.013	.001	-.001	.640

Table 2. % Bias β_1--Correctly Specified Model*

R^2	ρ_{12}**	Ungrouped	Randomly Grouped	Grouped by X_1	Grouped by X_2	Grouped by Y
.7	.75	-1.0%	-0.8	3.4	-2.2	38.1
.7	.5	-0.6	-0.5	-1.8	-0.2	34.5
.7	.25	-0.4	-0.4	0.0	1.3	37.9
.7	0	-0.3	-0.5	-0.3	-0.1	35.7
.3	.75	-2.2	-1.8	7.9	-5.2	150.9
.3	.5	-1.3	-1.2	-4.3	-5.5	165.9
.3	.25	-0.9	-1.0	0.0	3.1	165.2
.3	0	-0.6	-1.2	-0.6	-0.3	175.3

% Bias β_2--Correctly Specified Model**

R^2	ρ_{12}	Ungrouped	Randomly Grouped	Grouped by X_1	Grouped by X_2	Grouped by Y
.7	.75	1.3%	0.4	-4.5	2.2	37.3
.7	.5	0.9	0.1	3.5	0.6	40.9
.7	.25	0.7	-0.0	-6.4	0.2	37.0
.7	0	0.6	0.1	-5.2	0.6	39.3
.3	.75	3.0	1.0	-10.5	5.0	196.5
.3	.5	2.1	0.3	8.2	1.4	184.6
.3	.25	1.7	-0.1	-1.5	0.5	180.6
.3	0	1.4	0.3	-1.2	1.5	176.2

* Each entry calculated as follows: % Bias $(\hat{\beta}_1)$ = 100% times [Average Bias $(\hat{\beta}_1)/\beta_1$].

** The values of R^2 and of ρ_{12} given are those used to create the original 500-observation data sets.

*** Calculated same as % Bias $\hat{\beta}_1$.

Table 3. Relative Variance of Ungrouped to Grouped Estimators of β_1^*--Correctly Specified Model**

R^2	ρ_{12}	Random Grouping	Grouping by X_1	Grouping by X_2	Grouping by Y
.7	.75	.07**	.20	.09	.09
.7	.5	.08	.29	.08	.08
.7	.25	.08	.71	.08	.10
.7	0	.08	1.00	.09	.13
.3	.75	.07	.20	.09	.04
.3	.5	.08	.29	.08	.03
.3	.25	.08	.71	.08	.03
.3	0	.08	1.00	.09	.06

* Since X_1 and X_2 are created in the same way and $\beta_1 = \beta_2$, approximate results for the estimates of β_2 can also be obtained from this table. For example the relative variances for the estimators of β_2 under grouping by X_2 would be the entries in the column "Grouping by X_1," and so on.

** Each entry is the ratio of the average variance of the ungrouped estimator of β_1 to the average variance of the grouped estimator of β_1. (The higher the value of this fraction, the greater the efficiency of the grouped estimator relative to the ungrouped estimator.)

*** Figures have been rounded to the nearest hundredth of a percent.

Table 4. % Bias - Misspecified Model*

R^2	ρ_{12}**	Ungrouped	Randomly Grouped	Grouped by X_1	Grouped by X_2	Grouped by Y
.7	.75	75.1%	75.5	75.1	125.3	171.3
.7	.5	50.0	50.2	49.8	157.1	165.0
.7	.25	24.6	24.9	24.6	166.7	156.4
.7	0	-0.4	0.7	-0.4	-3.0	142.9
.3	.75	75.1	75.2	75.1	126.1	428.0
.3	.5	49.6	50.0	50.0	158.3	405.6
.3	.25	24.3	24.4	24.4	169.1	370.5
.3	0	-0.8	0.0	-0.8	-3.4	336.1

* Each entry calculated by the same method as in Table 1:
% Bias = 100% times [Average Bias $(\hat{\beta}_1(m))/\beta_1$].

** The values of R^2 and of ρ_{12} given are those used to create the original 500-observation data sets.

FOOTNOTES

1. Let: $\underline{y}$ be an (N by 1) vector containing the N observations for the dependent variable, $\underline{X}$ be an (N by k) matrix containing the observations for the k independent variables, $\underline{\beta}$ be a (k by 1) vector of regression coefficients, and $\underline{\varepsilon}$ be a (k by 1) vector of error terms.

2. The variance of the ε_i's is assumed to be constant across individuals, equal to var(ε).

3. Throughout this paper it is assumed that all m groups have exactly the same number of observations, n. (Thus, mn = N). Then $\underline{G} = [I_m \otimes \frac{1}{n} \ell_n']$, where I_m is an (m by m) identity matrix, $\otimes$ is a Kronecker product sysmbol, and ℓ_n is an (n by 1) column vector of 1's. We note in passing that when groups are of unequal sizes a generalized least squares treatment should be used to increase the efficiency of estimation. (cf. Feige & Watts 1972).

4. The normally distributed variates were generated by Marsaglia's rectangle-wedge method (cf. Knuth 1969).

5. This result can be found in Table 3, next to the last column, by use of the fact that X_1 and X_2 are generated symmetrically.

6. Feige and Watts (1972, pp. 346, 347, 351) appear to imply that merely grouping by values of the disturbance produces aggregation bias.

7. The situation in which grouping minimizes variation in confounding variables is discussed in Hannan (1976).

Bibliography

Blalock, Hubert M.
1964 *Causal Inferences in Nonexperimental Research.* Chapel Hill, N.C.: University of North Carolina Press.

Cramer, J.S.
1964 "Efficient grouping: regression and correlation in Engel curve analysis." *Journal of the American Statistical Association* 59 (March): 233-50.

Feige, Edgar L. and Harold W. Watts
1972 "An investigation of the consequences of partial aggregation of micro-economic data." *Econometrica* 40 (March): 343-60.

Hannan, Michael T.
1971 *Aggregation and Disaggregation in Sociology.* Lexington, Mass.: Heath.

Hannan, Michael T.
1976 "Aggregation gain reconsidered." Paper presented at annual meetings of the American Educational Research Association.

Hannan, Michael T. and Leigh Burstein
1974 "Estimation from grouped observations." *American Journal of Sociology* 39 (June): 374-92.

Hannan, Michael T.; Alice A. Young; and Francois Nielsen
1975 "Specification bias analysis of the effects of grouping of observations in multiple regression models." Paper presented at annual meetings of American Educational Research Association.

Hanushek, Eric A.; John A. Jackson, and John F. Kain
1974 "Model Specification, use of aggregate data, and the ecological correlation fallacy," *Political Methodology* 1 (August): 89-107.

Knuth, D.E.
1969 *The Art of Computer Programming*, vol. 2. Reading, Mass.: Addison-Wesley Publishing Co.

Johnston, J. *Econometric Methods*, 2nd edition, New York: McGraw-Hill.

Prais, S.J. and J. Aitchison
1954 "The grouping of observations in regression analysis." *Review of the International Statistical Institue* 22: 1-22.

Shively, W. Phillip
1969 "'Ecological' inference: the use of aggregate data to study individuals," *American Political Science Review* 63 (December): 1183-1196.

(DRAFT COPY)

THE PROBLEM OF DATA AGGREGATION AND CROSS-LEVEL INFERENCE WITH CATEGORICAL DATA

Carlyle E. Maw
National Institute of Education

Introduction

Educational issues are frequently expressed in terms of counts or enumerations. Questions such as how many students passed a course or about the number of students eligible for an educational program are typical questions posed by matters of educational policy or by research issues. Recent trends in educational measurement have produced new sources of enumeration data from mastery scores and from pass/fail classifications on criterion-referenced tests.

In contrast, the methodological focus of educational research and data analysis has been predominantly on measurement data, such as scales and test scores. There has been relatively less attention paid to methodological developments concerning enumeration data, in spite of the increasing prevalence of this type of data. Fortunately, much of the research into methodological problems involving measurement data can be generalized and applied to analytic problems with enumeration data. However, because of differing historical traditions in notation and in the formulation of analytic models, this generalization of solutions across types of data may not be obvious or straightforward.

The purpose of this paper is to generalize a methodological problem in educational research which is usually formulated with measurement data. The problem is perhaps best known as the "ecological fallacy," a term introduced from sociological research. Briefly, this fallacy refers to the difference in the correlation between two variables when the unit of analysis is changed from the group level to the individual level. For example, the correlation between test score and a measure of SES would in general be different when computed from classroom averages than when calculated from individual student data. This problem is of practical significance because often data is available at the group level but is either unavailable or more difficult to obtain at the individual level.

If the group-level correlation could be used to estimate the individual-level correlation, it would then be possible to perform an otherwise expensive or impossible analysis. Unfortunately, the cross-level inference can be made only under very restricted and not often realistic conditions. These restrictions will be discussed in detail later.

In general, this problem of cross-level inference exists for all measures of association between variables, from correlational measures to regression or asymmetric coefficients. In general, the problem is so severe that a coefficient at one level cannot be used to estimate the magnitude or even the sign of the coefficient at another level for analysis. In the context of educational research, the problem has been formalized and described in Burstein, 1974, 1975; Hannan, 1971; and Hannan and Burstein, 1974.

The cross-level inference problem just introduced for measurement data also arises with enumeration data. Reconsider, for example, the association between test score and SES. These measures can be transformed into classifications to give enumeration data. The individual association can then be measured between two dichotomies in a 2 X 2 contingency table with test scores and SES scales each categorized as "high" or "low". More generally, we could have more than two levels and instead of categorizing into dichotomies, we could classify into polytomies. This paper will focus on the more simple case of dichotomous classification.

To illustrate the problem of cross-level inference with enumeration data, consider the 2 X 2 table relating test score and SES for individual students. There are a variety of symmetric and unsymmetric measures of association that could be used to summarize this relationship (e.g. fourfold point correlation, chi square, and lambda). For the group level analysis, classrooms or schools could be classified as high or low with respect to average test score or average SES. Again some measure of association could be calculated for these aggregated dichotomies. As in the previous example which used measurement data, there is no general technique to use the coefficient calculated at one level of analysis to estimate either the magnitude or sign of the coefficient calculated at another level.

Several authors have attempted to qualify this general lack of relationship between coefficients at different levels of analysis and to develop specific sets of conditions or particular coefficients of association that would allow some cross-level inferences to be made. Some recent work with educational measurement data has shown that the introduction of a "grouping variable" (Burstein, 1974, 1975) provides a formalism which permits some prediction of the effect of grouping. In particular, the grouping variable notation focuses attention on the effect of different ways of assigning individuals to groups. This formalism provides a natural way to express and test the conditions that must be met in order to make cross-level inferences. The application of this idea to the context of enumeration data reveals a simplification that should be of practical use in problems of cross-level analysis. This will be described in detail below.

Summary of a Technique for Cross-Level Analysis With Measurement Data

This summary will address the problems of cross-level analysis in the context of a regression model with a single independent variable. This context has been the area of much study in economics and sociology (Blalock, 1961; Cramer, 1964; Prais and Aitchison, 1954; Theil, 1954; as well as the previously cited authors). Because it is the difference between coefficients calculated for two levels of analysis that is of interest, regression models will be developed for each level and then compared. Adopting the usual regression terminology with Y and X observed and e designating the unobserved error in the model, we have two models:

- individual level with n observations in the sample

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$\hat{y}_i = b_0 + b_1 x_i \qquad i = 1, 2, \ldots, n$$

- aggregate level with G observations in the sample

$$\bar{y}_j = \bar{\beta}_0 + \bar{\beta}_1 x_j + \bar{e}_j$$

$$\hat{\bar{y}}_j = \bar{b}_0 + \bar{b}_1 x_j \qquad j = 1, 2, \ldots, G$$

Where: Y, $\bar{Y}$, β are true values

$\hat{y}_i$, $\hat{\bar{y}}_j$, b are estimates

$\bar{y}_j = [\Sigma_i^{n_j} y_i]/n_j$, etc.

The question of practical interest for making cross-level inferences is whether there is bias in the estimate of β_1 using $\bar{b}_1$. That is, can the true individual-level regression coefficient be accurately estimated from the group-level regression coefficient observed on a sample of data. Burstein (1975) refers to the expected difference, $E(\bar{b}_1 - \beta_1)$ as the bias from grouping and denotes this bias as θ.

When groups are formed from individuals by randomly allocating individuals into aggregates, the value of θ should be close to zero. Blalock explains the reason for this lack of bias by noting that grouping can be expected to reduce the variability in X and in Y by about the same amount and therefore X will continue to account for about the same proportion of variation in Y as the level of analysis changes (Blalock, 1961, 104).

Although random assignment simplifies cross-level analyses, it rarely seems to be a realistic assumption to describe the allocation of students to classrooms or classrooms to schools. In fact, random allocation is almost certainly ruled out by the operation of social selection mechanisms which are very much a part of the educational or social system that is of interest.

Another mechanism for allocating individuals to groups might be on the basis of similar values of the independent variable X. In the above example, this would mean grouping students of relatively low SES together, students of higher SES together, and so on. Because schools tend to serve relatively homogeneous residential areas as a rule, and because SES is closely related to residential characteristics, it appears plausible to assume that students are roughly grouped according to SES characteristics in schools.

It has been observed by several authors that grouping by X allows cross-level inferences about regression coefficients (Blalock, 1961; Burstein, 1974, 1975; Cramer, 1964; Hannan, 1971; Prais and Aitchinson 1954). Because in many situations, such as the one just described for SES, grouping by X is not an unreasonable assumption, it is gratifying to discover that this situation also leads to a low or zero value for θ.

It cannot generally be concluded that grouping by X eliminates the ecological fallacy, however. Although cross-level inferences can be made with appropriate care for regression coefficients, cross-level inferences using other coefficients of association may be in error. The correlation coefficient, especially, changes rapidly as individuals are grouped by X. This is because grouping by X tends to preserve the variability in X across groups, since low X and high X individuals tend to remain separated by groups. This stability keeps the range of X relatively unchanged and the slope of the regression of Y on X then remains relatively constant (Blalock, 1961, 105). The variability in Y is reduced by grouping on X, however, unless X and Y are perfectly correlated. Since $bs_X = rs_Y$, r must increase when X and Y are not highly correlated. The correlation coefficient also may increase because errors of measurement tend to cancel in larger aggregates, and hence the correlation is less attenuated by error. The net result is that the correlation coefficient increases as individuals are grouped by X.

Although this summary is necessarily brief, it has introduced two points. First, grouping by X preserves the regression slope so that some cross-level influences can be made with little expected bias (0). Second, different coefficients of association behave differently even under the same context of grouping, so that conditions which allow some cross-level inferences for one coefficient of association may not permit using another coefficient.

Summary of a Technique for Cross-Level Analysis with Categorical Data

Perhaps the best-known reference on the ecological fallacy is Robinson's (1950) paper, although the problem was discussed earlier in the context of education by Walker (1928). Robinson used a 2 X 2 table as his example and noted that for the individual-level association, the fourfold point correlation uses the internal cell frequencies. He then noted that correlations calculated at the group level use only the marginal frequencies of the group tables and ignore the internal cells in computing the "ecological" or group-level correlation. Because, as Robinson claims, "...it is well known that the marginal frequencies of a fourfold table do not determine the internal frequencies" (Robinson, 1950, 354), the information content of the two correlation coefficients is different. That is, the ecological correlation coefficients use only part of the information in the data and exclude another part that is used to calculate the individual-level correlation coefficient. Robinson then notes the large differences in numeric values of the resulting correlation coefficients at different levels of analysis and concludes that "...there need be no correspondence between the individual correlation and the ecological correlation" (Robinson, 1950, 354).

Although this conclusion would seem to preclude any cross-level inference using fourfold tables, there have been attempts to establish some connection between coefficients of association calculated at one level with those calculated at a different level of analysis. Two approaches will be discussed, one by Duncan and Davis (1953); the other by Goodman (1953, 1959).

The first approach uses the fact that the marginal frequencies of a fourfold table do at least establish bounds for the variation of the interior cell values. Given a set of marginals for a fourfold table, there are some internal cell values which are clearly impossible, and, hence, the range of internal variation is limited. This corresponds to the case (to be described below) where the group-level tables establish the individual-level table marginals and hence limit the possible values of the interior cells for the individual-level table. With this limitation on the interior, there is a corresponding limit on the range of the fourfold point correlation at the individual level because this coefficient is calculated from the internal cell values. Unfortunately the range restriction is rather broad and even allows sign changes, so that the practical use of this technique for cross-level inference has been limited.

The second approach by Goodman exploits the fact that there are usually many group-level tables which can be treated as subtables of the individual-level table. That is, the individual-level fourfold table can be thought of as broken down into subtables when the individual data is grouped by classroom or by school. Goodman proposed a linear model to use not just the marginal information from the group-level tables as in the first approach but in addition to incorporate information about the distribution of marginal values across the set of group-level tables.

More specifically, by making some assumptions about the distribution of internal frequencies across the subtables, the information from these observed marginals can be used to predict the internal frequencies of the individual level table. It is then an obvious step to compute the measure of association between X and Y for the individual-level data.

Before summarizing Goodman's argument, some notation is in order. Figure 1 illustrates a fourfold table where Y could be test score and X a measure of SES.

		X		
		0	1	
Y	0	f(11)	f(12)	f(1+)
	1	f(21)	F(22)	f(2+)
		f(+1)	f(+2)	n

Figure 1. Cross-classification of X and Y showing cell frequencies in total sample of n students at individual level of analysis.

By grouping individual students into classrooms or schools, the table in Figure 1 can be thought of as being broken down into separate subtables, one for each classroom or school. If Z stands for a random variable which takes on values z = 1, z = 2, ..., z = G, for G groups, then Z represents a counter for the various classrooms or schools. Figure 2 illustrates the use of Z as a counter with appropriate notation.

		X		
		0	1	
Y	0	f(111)	f(121)	f(1+1)
	1	f(211)	f(221)	f(2+1)
		f(+11)	f(+21)	n(1)

Z = 0

		X		
		0	1	
Y	0	f(112)	f(122)	f(1+2)
	1	f(212)	f(222)	f(2+2)
		f(+12)	f(+22)	n(2)

Z = 1

where: f(111) + f(112) = f(11+) = f(11), as in Figure 1
n(1)+ n(2) = n(+) = n, as in Figure 1

Figure 2. Cross-classification of X and Y for two group-level subtables.

By totalling the marginals of the subtables in Figure 2, we can establish the marginals of the individual-level table of Figure 1. But, as Robinson (1950) has noted, this is not enough information to let us fix the values of the internal cell entries of the table in Figure 1. Goodman's suggestion was to form a linear model which used the marginal proportions of the group-level tables to predict the internal cell values of the individual-level table.

The linear model for this purpose will again be the simple regression model with a single independent variable, as described before. In this case, however, the definitions of X and Y will be the relative proportions of the X and Y variables at the low value of their dichotomized values. These new transformations of X and Y will be denoted by lower case x and y. Referring to Figure 2, the values of x would be: x_1; x_2= f(+11)/n(1), f(+12)/n(2); and the values of y would be: y_1, Y_2 = f(1+1)/n(1), f(1+2)/n(2). In a real situation, there would be G (x,y) pairs of values from the G subtables.

The model proposed by Goodman is:

$$Y = r + (p - r)\, X + e \tag{1}$$

where: X, Y are random variables with values

$x_1, y_1; x_2, y_2; \ldots; x_G, y_G$ just defined

$r = \text{Prob}\,(Y = 0 \mid X = 1) = f(12)/f(+2)$

$p = \text{Prob}\,(Y = 0 \mid X = 0) = f(11)/f(+1)$

The numerator of each X is a binomially distributed random variable with parameters $n(Z)$ and $\pi(+1Z)$ for each subtable, $Z = 1, \ldots, G$. Similarly, the interior cells $f(1lZ)$ and $f(12Z)$ of all subtables are also binomially distributed with parameters $f(+1Z)$ and $f(+2Z)$ for respective sample sizes and parameters $\pi(11)$ and $\pi(12)$ for respective conditional probabilities. These are conditional probabilities given $X = 0$ and $X = 1$, respectively and are assumed to be constant over all subtables (hence independent of Z). With these definitions, it now follows that the numerator of Y can be defined as the sum of two conditional probabilities multiplied by their sample sizes (Iverson, 1969).

$$f(1+Z) = \pi(11)\, f(+1Z) + \pi(12)\, f(+2Z) \tag{2}$$

Taking expectations with respect to Z:

$$E[f(1+Z)] = \pi(11) E[f(+1Z)] + \pi(12)\, E[f(+2Z)]. \tag{3}$$

If we now assume that the column marginals are fixed, these expectations on the right-hand side of (3) become the sample values and, hence:

$$E[f(1+Z)] = \pi(11)\, f(+1Z) + \pi(12) f(+2Z), \tag{4}$$

which is recognized as the same as Goodman's model by rearranging terms and dividing by $n(Z)$. Using the definition for x and y above equation (1):

$$\begin{aligned} E[y \mid x] &= \pi(11)X + \pi(12)(1-X) \\ &= \pi(12) + (\pi(11) - \pi(12))X, \end{aligned} \tag{5}$$

which is Goodman's model of equation (1) with $p = \pi(11)$ and $r = \pi(12)$, as defined above equation (2) and depicted in Figure 1. Because the columns are fixed, these expectations are understood to be conditional on x, so that the left-hand side of (5) is written $E(y|x)$.

This completes the technique of cross-level inference, for $\pi(11)$ and $\pi(12)$ are conditional probabilities defined for the individual-level table in Figure 1. Because the individual marginals are fixed by the observed subtable marginals, either $\pi(11)$ or $\pi(12)$ alone is sufficient to fix all of the cells of the individual level table. From there it is a simple calculation to compute the fourfold point correlation coefficient, as in the Robinson (1950) example.

Assumptions in the Model

As noted in the derivation of this model (5), there are several assumptions that need to be met. In addition to the usual distributional assumptions for regression models, there is also the requirement that $\pi(11Z) = \pi(11)$ and $\pi(12Z) = \pi(12)$ for all Z. This requirement was relaxed in a later paper (Goodman, 1959) to the conditions that:

$E[\pi(11Z)|x] = K_1$ and $E[\pi(12Z)|x] = K_2$. These conditions mean that the expected values of the conditional probabilities denoted by Goodman as p and r must be constant across subtables. Whether or not these restrictions are often met in practical situations has been a subject of debate. What is clear from the literature is that Goodman's technique for cross-level inference has not received much practical application and some authors who have noted this attribute it to the difficulty in meeting these conditions or to the risk of badly biased estimates if the assumptions are not met (Cartwright, 1969; Shively, 1969).

The assumptions are unfortunately stated as conditions involving unobserved quantities, so that there is some difficulty in formulating a rule-of-thumb for deciding when to apply this technique. In order to develop one such rule-of-thumb, two examples will now be presented. The first example uses simulated data which has been grouped into categories in specified ways to illustrate the effect of grouping in a simple case involving only two groups. The second example will use another simple example but with real data to illustrate some practical problems in applying the rule-of-thumb.

An Illustrative Example with its Geometric Representation

This example is intended to show the effect of grouping on the assumptions that underlie the regression technique for cross-level inference as proposed by Goodman. The individual-level table shown in Figure 3 is divided into two "group-level" subtables. This grouping process is performed twice: once according to a grouping rule A that assigns individuals to the two subtables according to their Y values. The example does not depict an extreme case of grouping by Y but rather a strong tendency for such grouping. Grouping rule B results in two other subtables also shown in Figure 3 which illustrate a tendency to group by values of X (high X students in the first subtable (I); low X predominate in the second (subtable (II)*.

Individual level table

		X 0	X 1	
Y	0	100	50	150
	1	80	70	150
		180	120	300

Grouping Method A

80	20	100
40	10	50
120	30	150

Subtable I

20	30	50
40	60	100
60	90	150

Subtable II

---

* The technical rule for grouping Method A was to assign 67% of the 150 low-Y individuals to the low-Y row in Subtable I. Then 67% of the high-Y individuals were assigned to the high-Y row of Subtable II. Grouping Method B randomly allocated high- and low-Y individuals to the two subtables but kept the same X marginals as grouping Method A. Hence, Method B could be referred to as grouping by X.

Grouping Method B

Subtable I

67	13	80
53	17	70
120	30	150

Subtable II

33	37	70
27	53	80
60	90	150

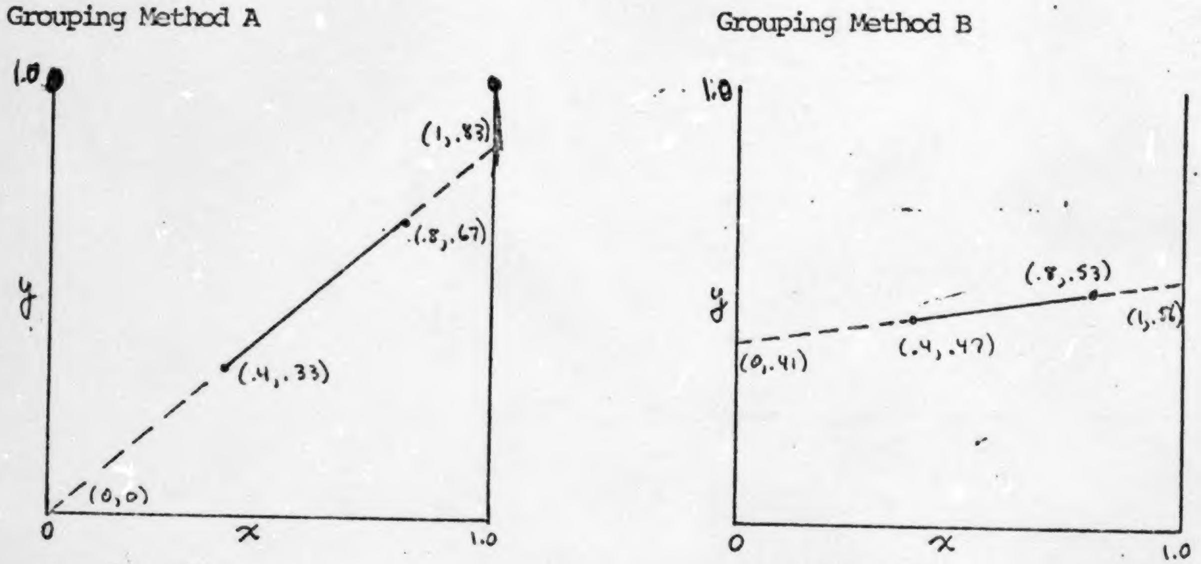
Figure 3. Exemplary grouping Methods A and B with resulting subtables from the individual-level table.

By referring to Table 1, we see that the assumptions of constant $\pi(11Z)$ and $\pi(12Z)$ have been violated for grouping Method A but that these assumptions have been met by grouping Method B. More extreme grouping by X or Y would yield consistent but more exaggerated patterns. A graphical illustration will show why this is so.

	Individual-level table	Grouping Method A		Grouping Method B	
		Subtable I	Subtable II	Subtable I	Subtable I
$p = \pi(11)$	.56	.67	.33	.56	.55
$r = \pi(12)$	.42	.67	.33	.43	.41
x	.60	.80	.40	.80	.40
y	.50	.67	.33	.53	.47

Table 1. Comparison of p and r values for two methods of grouping individuals into aggregates.

Using the values shown in Table 1 for x and y, we can graph the line described by the regression model (5). This graph of the two (x,y) points associated with the two subtables can be drawn for grouping Methods A and B and the resulting regression lines are shown in Figure 4. Note that because there are only two groups and hence two (x, y) points, the regression line need not be estimated in this case.



Grouping Method A

Grouping Method B

$y = 0 + .83x$
$p = .83$
$r = 0$

$y = .41 + .15x$
$p = .56$
$r = .41$

Figure 4. Regression lines for two grouping methods extrapolated to extreme values to estimate parameters p and r.

When the regression lines of Figure 4 are extrapolated to the extremes of $x = 0$ or $x = 1.0$, the results have an interesting interpretation. When $x = 0$, the extrapolation in effect places all of the data in the second column ($X = 1$) so that the conditional probability r is the same as the marginal value of y. Similarly, when $x = 1.0$, all of the data is placed in the first column and the conditional probability p is the same as y. For grouping Method A, the extrapolation does a poor job of predicting the individual-level table values and the bias from violated assumptions is clear. On the other hand, grouping Method B does very well in predicting the values for p and r.

The graphical representation in Figure 4 shows why p and r must be nearly constant for given values of x in order to avoid bias from extrapolation.

Although there are a number of grouping techniques that should be simulated to give a convincing argument, instead let us turn to a simple example with real data.

Illustrative Example with Real Data

The data for this example comes from the evaluation of the Emergency School Assistance Program (ESAP) carried out at the National Opinion Research Center (Crain, et al, 1973). Only the fifth grade sample of Southern schools will be used, but that data includes 20,157 students in 758 classrooms in 343 schools. (This is somewhat less than full sample because of excluding classrooms of less than 17 students or greater than 37).

The Y factor for the cross-classification and regression model will be student test scores on a standardized achievement test, dichotomized at above and below the sample mean. Two X factors will be chosen to correspond to grouping methods A and B of the previous example. For the first X factor, a short attitude scale about self-image will be used and again dichotomized above and below the mean. This factor will be called xself. Since attitudes about oneself should vary individually and not be clustered by school, this factor should resemble grouping Method A.

For the comparison X factor, we will use a short scale which measures attitudes about the student's school. Such a factor should group students by schools, since the school itself is the referrent for the attitude questions. To the extent that students share their attitudes about the school they attend, this X factor, called xschool, should result in aggregating students like grouping Method B. Of course, students are not literally grouped into a school on the basis of their attitudes about their school, but the assumed similarity of attitudes among students in the same school will result in data that will be similar to data that would have been observed if school attitude had indeed been such a grouping factor. Thus, xschool becomes a good example of what would happen if we could experiment with grouping instead of just observing naturally-formed groups.

In the ESAP data, data exists at the individual student level, the classroom level, and the school level. Therefore, there is the possibility of making cross-level inferences analyses with the categorical variables created from test scores and attitude scales and of then verifying the cross-level inferences by comparing the results at the individual level of analysis.

The components of the two X factors are listed in Table 2. The scales are simple additive scales of those question responses which indicate positive attitudes receiving a score of one and zero otherwise. The resulting scales are dichotomized at the mean and scored zero/one for individuals, classrooms, and schools.

Attitude questions about oneself

1. Do you think you are better than more students at doing school work, about the same, or not as good as most students?
2. Are you satisfied with yourself?
3. Do you think you can do things as well as most students can?
4. When you make plans, are you almost sure you can make them work?
5. Would you say you are very happy, pretty happy, or not too happy these days?
6. At school, are you often blamed for things that just aren't your fault?
7. Do you think it doesn't pay to try hard because things don't turn out right anyway?

Attitude about school

1. Is there any adult at this school you could talk to if you were upset or in trouble?
2. Do you think your teacher likes you?
3. In the past week, did you think any of your school work or homework was fun?
4. Do you usually hate school?
5. Do you like your teacher?
6. Is reading too hard for you now?
7. Is arithmetic too hard for you now?
8. Do you think most of the rules in your classroom are fair?
9. Do you get really angry when teachers try to make you do things you don't want to do?
10. Do you feel like you don't really belong in this school?

Table 2. Items in X factor scales.

Table 3 presents a summary of the findings from the ESAP analysis at the individual level and using the regression method of cross-level inference. As expected, the predicted values of p and r are much closer to the individual values for xschool than for xself.

Individual level	p	r	p-r
score x xschool	.567	.401	.166
score x xself	.576	.377	.199
Classroom level			
score x xschool	.556	.414	142
score x xself	.811	.195	.612
School level			
score x xschool	.510	.455	.055
score x xself	.918	.144	.764

Table 3. Calculated and estimated values of p and r from the ESAP data at different levels of analysis

Although the example has confirmed the prediction that grouping by X allows some forms of cross-level inference, one step remains to complete the generalization of this finding from previous work with measurement data. The parameters p and r should be formally identified with regression coefficients, since it is only for these coefficients that the claim of stability is made across different levels of analysis.

To make this analogy, note that the regression coefficient b can be calculated from:

$$b = \frac{\Sigma XY - (\Sigma X)(\Sigma Y)/n}{\Sigma X^2 - (\Sigma X)^2/n} \qquad (6)$$

When X and Y are coded 0, 1 as in Figure 1. By substituting the following quantities:

$$\Sigma X = \Sigma X^2 = f(+2)$$

$$\Sigma Y = \Sigma Y^2 = f(2+),$$

$$b = \frac{f(22)n - f(+2)\ f(2+)}{f(+2)n - f(+2)\ f(+2)}$$

$$= \frac{f(+1)f(22) + f(+2)\ f(22) - f(+2)\ f(2+)}{f(+2)f(+1) + f(+2)\ f(+2) - f(+2)\ f(+2)}$$

$$= \frac{f(11)\ f(22) - f(12)\ f(21)}{f(+2)\ f(+1)}$$

$$= \frac{f(11)\ f(+2) - f(12)\ f(+1)}{f(+2)\ f(+1)}$$

$$= \frac{f(11)}{f(+1)} - \frac{f(12)}{f(+2)}$$

$$= p - r \quad . \text{ QED}$$

Thus, the quantity (p − r) is formally identical to the regression coefficient for dichotomous data and the results established for measurement data appear to hold for qualitative data.

Although grouping by X appears to stablize the cross-level inferences for p and r from the group level to the individual level, the variances of p and r calculated across the classroom and school groups are not greatly affected by the choice of X factor. That is, the differences in grouping factors does not in this particular example appear to affect the stability of p and r across subtables. Table 4 gives the calculated values of mean p and r for the classroom and school groups using both X factors. The expected stabilization of p and r with xschool is not observed, even though the predicted values of p and r from this X factor show little bias.

Classroom level	$\bar{p}$	std. dev.	$\bar{r}$	std. dev.
xschool	.574	.262	.403	.250
xself	.564	.236	.396	.255
			n = 758	
School level				
xschool	.580	.204	.409	.208
xself	.573	.185	.399	.212
			n = 343	

Table 4. Mean values and standard deviations of p and r calculated at classroom and school levels.

This table is not in accord with Goodman's prediction because the relative variability of $\bar{p}$ is less for xself than for xschool in all cases. Hence, one might expect a greater bias for the regression estimate of p using xschool than for xself. Clearly, more needs to be known about the data than just the expected stability of p in order to risk a cross-level inference.

Conclusion

As in the case with measurement data, the conclusion is that grouping by X allows some cross-level inferences to be made, albeit with caution. The appropriate coefficients again appears to be the regression coefficients, which in case of the categorical data are f(11Z)/f(+1Z) or p and f(12Z)/f(+2Z) or r. In fact, the quantity (p - r) is identical to the regression slope computed for a fourfold table.

References

Blalock, H. M. Causal Inferences in Nonexperimental Research. New York: Norton, 1961.

Burstein, L. Issues Concerning Inferences From Grouped Observations. Paper presented at the Annual Meeting of the American Educational Research Association, 1974. Chicago, Illinois.

Burstein, L. Data Aggregation in Educational Research: Applications. Paper presented at the Annual Meeting of the American Educational Research Association, 1975, Washington, D. C.

Cartwright, D. Ecological Variables. In E. Borgatta (Ed.), Sociology Methodology. San Francisco, Calif.: Jossey-Bass, 1969, 155-218.

Crain, R., et al, Southern Schools: An Evaluation of the Effects of the Emerging School Assistance Program and of School Desegregation (2 Vols.). Chicago: National Opinion Research Center, 1973.

Duncan, O. D. and Davis, B. - An Alternative to Ecological Correlation. American Sociological Review, 1953, 18, 665-666.

Goodman, L. A. Ecological Regressions and the Behavior of Individuals. American Sociological Review, 1953, 18, 663-664.

Hannan, Michael T. Aggregation and Disaggregation in Sociology. Lexington, Mass.: Heath, 1971.

Hannan, Michael T. and Burstein, L. Estimation From Grouped Observations. American Sociological Review, 1974, 39, 374-392.

Hannan, Michael T., Freeman, J. H., and Meyer, J. W. Specification of Models for Organizational Effectiveness. American Sociological Review, 1976, 41, 136-143.

Iversen, G. R. Estimation of All Entries in Contingency Tables When Only Margins are Observed. Unpublished doctoral dissertation, Harvard University, 1969.

Prais, S. J. and Aitchison, J. The Grouping of Observations in Regression Analysis. Review of the International Statistical Institute, 1954, 22, 1-22.

Robinson, W. S. Ecological Correlations and The Behavior of Individuals. American Sociological Review, 1950, 15, 351-357.

Shively, W. P. "Ecological" Inference: The Use of Aggregate Data To Study Individuals. American Political Science Review, 1969, 63, 1183-96.

Walker, H. M. A Note on the Correlation of Averages. Journal of Educational Psychology, 1928, 19, 636-642.